International Journal of Engineering Technology and Management Sciences
Website: ijetms.in Issue: 3 Volume No.8May - June – 2024
DOI:10.46647/ijetms.2024.v08i03.020 ISSN: 2581-4621

Deep Learning-Based Feature Extraction for Speech Emotion Recognition

[1]Dharmendra Kumar Roy, [2]Naga Venkata Gopi Kumbha, [3]Harender Sankhla,

[4]G. Teja Alex Raj, [5]Bashetty Akhilesh

[1]Associate Professor, Department of Computer Science and Engineering, Hyderabad Institute of Technology and Management, Hyderabad, India
[2]Department of Computer Science and Engineering(AI&ML), Hyderabad Institute of Technology and Management, Hyderabad, India
[3]Department of Computer Science and Engineering(AI&ML), Hyderabad Institute of Technology and Management, Hyderabad, India
[4]Department of Computer Science and Engineering(AI&ML), Hyderabad Institute of Technology and Management, Hyderabad, India
[5]Department of Computer Science and Engineering(AI&ML), Hyderabad Institute of Technology and Management, Hyderabad, India

***Abstract*--** Emotion recognition from speech signals is an important and challenging component of Human-Computer Interaction. In the field of speech emotion recognition (SER), many techniques have been utilized to extract emotions from speech signals, including many well-established speech analysis and classification techniques. This model can be built by using various methods such as RNN, SVM, deep learning, cepstral coefficients, and various other methods, out of which SVM normally gives us the highest accuracy. We propose a model that can identify emotions present in the speech, which can be identified by various parameters such as pitch, speaking rate, speech time, and frequency patterns. Emotion detection in digitized speech contains 3 components: Signal processing, Feature extraction, and Classification. The model first tries to remove the background noises then extract the features present in the speech and classify it into a single emotion. This model is capable of identifying seven different emotions that can be found in human speech. We can use different classifiers like GMM and HMM to classify features such as Spectral Subtraction, Wiener Filtering, Adaptive Filtering, and Deep Learning Techniques. This model can be used in various fields such as healthcare, security, psychology, medicine, education, and entertainment.

Keywords-– Speech Emotion Recognition, Emotion Detection, Speech Signal Processing, Feature Extraction, Classification, Deep Learning, Seven Basic Emotions, Machine Learning.

I.INTRODUCTION

Emotion detection from speech is a relatively new field of research with exciting potential applications. In human-computer or human-human interaction systems, emotion recognition systems could provide users with improved services by adapting to their emotions. In virtual worlds, emotion recognition could help simulate more realistic avatar interactions. The body of work on detecting emotion in speech is still limited. Researchers are currently debating which features influence the recognition of emotion in speech. There is also considerable uncertainty regarding the best algorithm for classifying emotions and how to group different emotions. For a machine to understand the

mindset/mood of humans through a conversation, it first needs a solid foundation. Implementing a speaker and speech recognition system is essential, as it allows the machine to identify the participants and their spoken words. Speech analysis can then be performed on the data extracted from these processes. Accurately identifying human emotions unlocks a range of valuable applications. For instance, computers can perceive and respond to human non-lexical communication, such as emotions. In such cases, after detecting human emotions, machines could customize settings based on individual needs and preferences. Recent advancements [1] have enabled the description of images and videos using a single high-level sentence, thanks to the availability of large datasets [2, 3].

A variety of temporal and spectral features can be extracted from human speech. Mel Frequency Cepstral Coefficients (MFCCs) and pitch serve as solutions to this issue, along with Formants of Speech as inputs to classification algorithms. The emotion recognition accuracy of these experiments reveals which features contain the most emotional information and the rationale behind it. It facilitates the development of criteria for grouping emotions together. High emotion recognition accuracy can be attained through these techniques.

The idea emerged during the execution of our hobby project, 'Emotion Detection using Image Processing. Despite efforts to enhance accuracy by increasing features and dataset quality, significant improvement remained elusive. This led to the proposition of integrating speech and image features to potentially enhance model accuracy. Internet research revealed that this area is currently under-researched. Additionally, inspiration was drawn from Google's Duplex call video, motivating the pursuit of this project.

II.LITERATURE SURVEY

Considering the project, research, and successful projects and papers were sought over the internet and the IEEE official website. Surveys by Maisy Wieman and Andy Sun provided accurate information about vocal pattern analysis for emotion detection.[6]

Algorithms and methods used for vocal pattern feature determination, such as the MFCC algorithm, were identified. Recent advancements in deep learning, as discussed by Arjovsky et al. (2017)[1], allow the description of images and videos using a single high-level sentence, facilitated by large datasets described in works like those by Lin et al[2]. (2014) and Krishna et al. (2017)[3].

The paper "Speech-Based Human Emotion Recognition Using MFCC" by M.S. Likitha et al. discusses emotion extraction from speech using MFCC. The process involves converting MP3 files to WAV format, MFCC feature extraction (including Frame Blocking, Windowing, Fast Fourier Transform, Mel Frequency Warping, and MFCC calculation), and decision-making based on mean and standard deviation. This approach enables emotion identification in audio clips with potential applications in human-computer interaction and affective computing.[7]

"Speech Emotion Recognition Using Deep Learning Techniques: A Review" by Ruhul Amin Khalil et al. provides insights into two approaches for building Speech Emotion Recognition (SER) systems: the traditional approach and deep learning techniques.[4] Traditional methods involve steps like Speech Preprocessing, Feature Extraction Selection, Classification, and Recognition, while deep learning techniques offer alternative methods like Deep Boltzmann Machine (DBM), Recurrent Neural Network (RNN), Recursive Neural Network (NN), Deep Belief Network (DBN), Convolutional Neural Network (CNN), and Auto Encoder (AE). The paper comprehensively reviews deep learning techniques for SER, detailing their architectures and applications in classifying various natural emotions.

In recent advancements, the ATT-LSTM model has been proposed, combining the attention mechanism with LSTM to enhance multivariate time series analysis. This integration allows the LSTM to filter out irrelevant information and effectively capture interactions within the sequences, improving its predictive capabilities. Experiments conducted on two real-world datasets, Nasdaq 100 and Beijing PM2.5, demonstrate the model's superior performance compared to six other models, as

International Journal of Engineering Technology and Management Sciences
Website: ijetms.in Issue: 3 Volume No.8May - June – 2024
DOI:10.46647/ijetms.2024.v08i03.020 ISSN: 2581-4621

evaluated by MAE and RMSE metrics. Additionally, predictions made by the ATT-LSTM model closely align with actual data, further demonstrating its effectiveness and robustness in handling complex time series data.[8]

"Enhanced Speech Emotion Recognition Through Combined Acoustic and Linguistic Features" by Misaki Sakurai and Tetsuo Kosaka proposes a method to improve SER by integrating acoustic and linguistic features.[5] The acoustic model (AM) is trained using emotional speech data, while the language model (LM) is adapted with a mixed n-gram method to understand emotional context better. Linguistic features are extracted using the eVector method, and acoustic features are obtained using deep neural networks. Finally, linguistic and acoustic scores are combined to generate a comprehensive emotion score, promising enhanced emotion recognition accuracy for applications in human-computer interaction and affective computing.[5]

In recent studies, various baseline models incorporating LSTM, BLSTM, and attention mechanisms were compared to assess performance on speech data. The proposed method demonstrated superior mean and maximum accuracy compared to these baselines. Experimentation revealed that optimal performance was achieved with a minimum duration of 10 ms and a 0.01% threshold for silence removal. For consistency and generalizability, it is suggested that this method be tested on additional datasets and languages in future research.[9]

III.METHODOLOGY

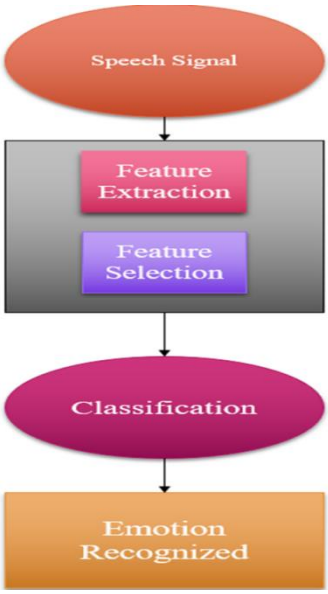

Fig. 1

Speech Signal Dataset:

Referral was made to two datasets, namely RAVDESS and SAVEE Dataset, specifically extracting audio data. The RAVDESS database, comprising 24 professional actors, contains expressions of calm, happy, sad, anger, fear, surprise, and disgust, each produced at two emotional intensity levels, along with a neutral expression. From this, 2000 audio samples in WAV format were obtained. The SAVEE dataset, recorded from 4 male actors, includes 7 emotions with 15 sentences each. A customized dataset was formed by merging these two datasets, ensuring a balance in emotion representation across 7 folders, each housing specific emotional samples.

International Journal of Engineering Technology and Management Sciences
Website: ijetms.in Issue: 3 Volume No.8May - June – 2024
DOI:10.46647/ijetms.2024.v08i03.020 ISSN: 2581-4621

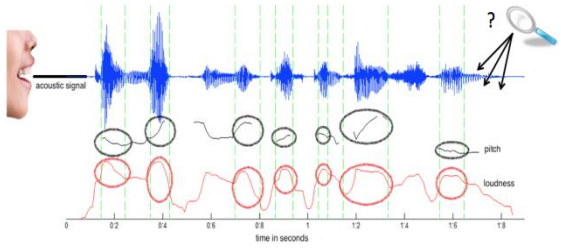

Fig. 2

Feature Extraction:

The complexity of time-domain sound representation necessitates the mapping of this representation into informative features. Techniques involve assessing signal energy, duration, statistics, and Fourier Transform for frequency domain processing. Mel-frequency Cepstrum Coefficients (MFCC) capture frequency characteristics, aligning well with human hearing perception. Pitch variations over time are measured on coarse and fine scales to record pitch changes.

Feature Selection:

Due to a high variance in extracted features, a filtering process was undertaken to determine the most significant ones. Approximately 72 windows per audio sample were windowed, resulting in a total of 577 features per windowed sample and 41,558 features in total. Heuristics were employed to score each feature for importance.

CNN AND LSTM:

A basic convolutional network comprises convolutional, pooling, and output layers. Convolution layers extract features while pooling layers reduce parameters. Convolutional and pooling layers are added successively to extract specific features. The output layer transforms the output into desired classes. Long Short-Term Memory Networks (LSTMs), a type of RNN, excel at learning long-term dependencies and consist of four interacting neural network layers.

LSTM NETWORKS:

LSTMs address the long-term dependency problem by design, featuring a chain-like structure with four interacting neural network layers. Each line in the diagram represents a vector, with pink circles representing pointwise operations and yellow boxes signifying learned neural network layers. Lines merging denote concatenation while forking lines indicate content duplication.

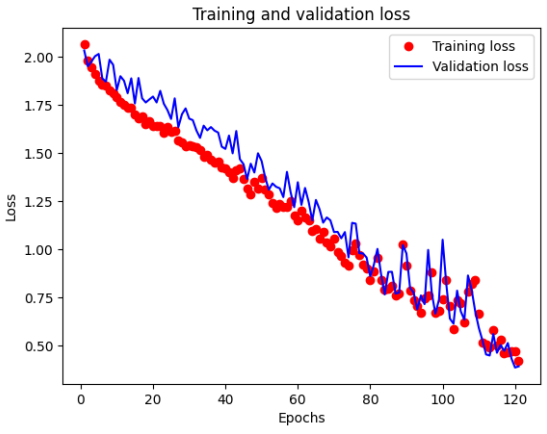

Fig. 3

The field of Speech Emotion Recognition (SER) stands ripe for advancement, with opportunities to innovate upon traditional approaches. In the initial stage of data preprocessing, integrating various techniques like noise removal, data augmentation, and addressing overfitting concerns can substantially enhance model performance.

Noise removal methods, such as spectral subtraction and deep learning-based denoising algorithms, can improve signal clarity, crucial for accurate emotion recognition amidst diverse environmental noises.

Augmenting the dataset with techniques like time stretching and pitch shifting not only enriches the training data but also aids in adapting models to diverse speaking styles and emotional

International Journal of Engineering Technology and Management Sciences
Website: ijetms.in Issue: 3 Volume No.8May - June – 2024
DOI:10.46647/ijetms.2024.v08i03.020 ISSN: 2581-4621

expressions.

Mitigating overfitting through regularisation techniques and exploring ensemble learning strategies further fortify SER models against memorizing training data, facilitating better generalization to real-world scenarios.

During the feature extraction stage, it is possible to extract several features that provide multiple insights from the data. There are various features that can be extracted for tasks like emotion recognition, speech recognition, and music information retrieval. Some common audio features include

Spectral features: These are based on the magnitude of the Fourier Transform (STFT, Mel-spectrogram, MFCC, etc.) and capture information about the distribution of energy in frequency bands.

Temporal features: These focus on changes in the audio signal over time.

Statistical feature: These represent certain statistical properties of the audio signal.

Cepstral features: These describe the spectral envelope in terms of its periodic structure and are often used as a compact representation of spectral features.

Harmonic features: These represent the harmonic structure of a signal.

Filter bank features: These represent the distribution of energy in different frequency bands.

Building a hybrid model can rapidly increase the model's efficiency and performance rates.

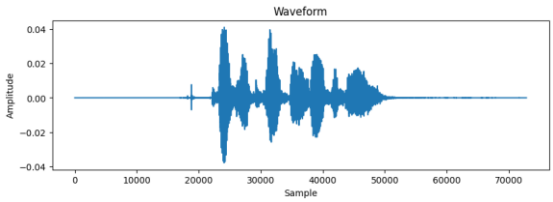

Fig. 4

We can consider combining different deep learning architectures, one approach is to use a multi-stream architecture, where each stream consists of a different deep learning model. The outputs of each stream can then be combined, using a technique such as concatenation or fusion, to make the final emotion prediction.

Another approach is to use a hierarchical architecture, where the output of one deep learning model is fed as input to another deep learning model. For example, the the output of a CNN can be fed as input to an RNN or LSTM, allowing the latter model to capture both local patterns and temporal dependencies in the audio signal.

IV.RESULT AND DISCUSSION

Finally, we have achieved 97% accuracy in predicting the seen data but the model is still imperfect in performing well on unseen data.

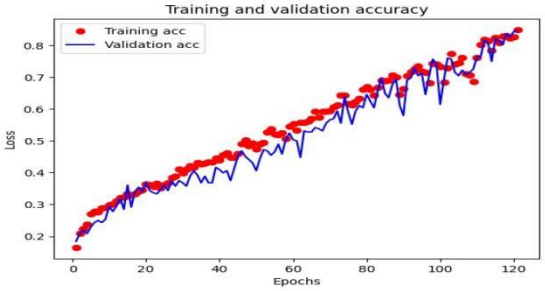

Fig. 5

In conclusion, this project has successfully classified different emotions that are usually found in human speech. Though the accuracy is around 97 on a trained dataset when performed in live demonstrations the classification of these emotions is imperfect. This issue can be resolved by providing the model with a more efficient dataset, this will increase the performance and accuracy

International Journal of Engineering Technology and Management Sciences
Website: ijetms.in Issue: 3 Volume No.8May - June – 2024
DOI:10.46647/ijetms.2024.v08i03.020 ISSN: 2581-4621

even in live situations. To use this model in live situations we have provided an audio recorder with the model which can record the audio for a predetermined period.

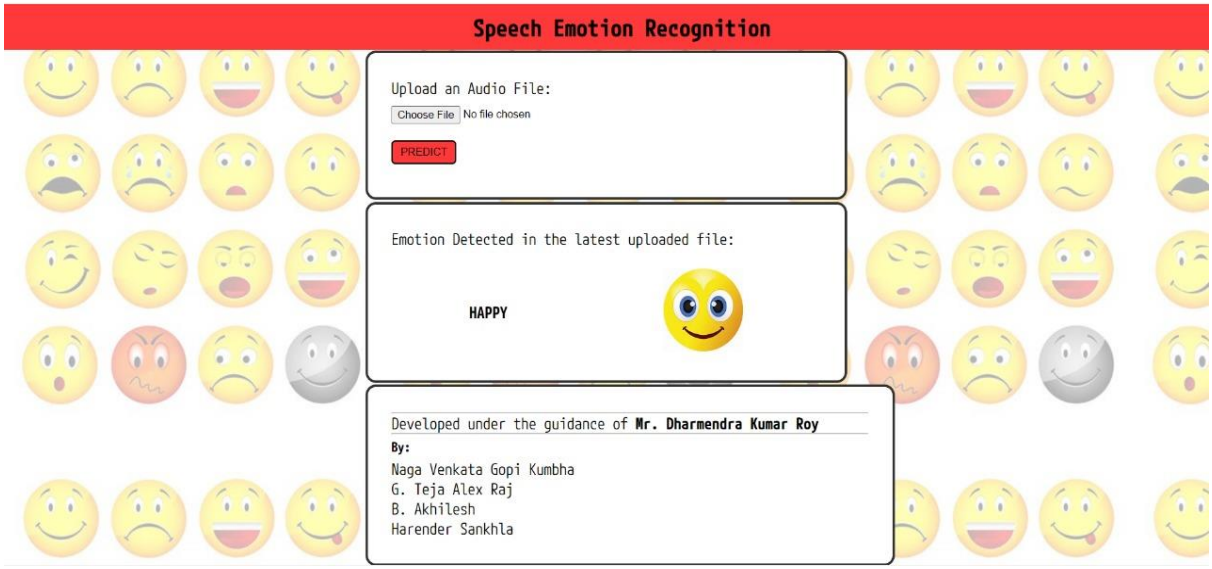

Fig. 6

Fig. 7

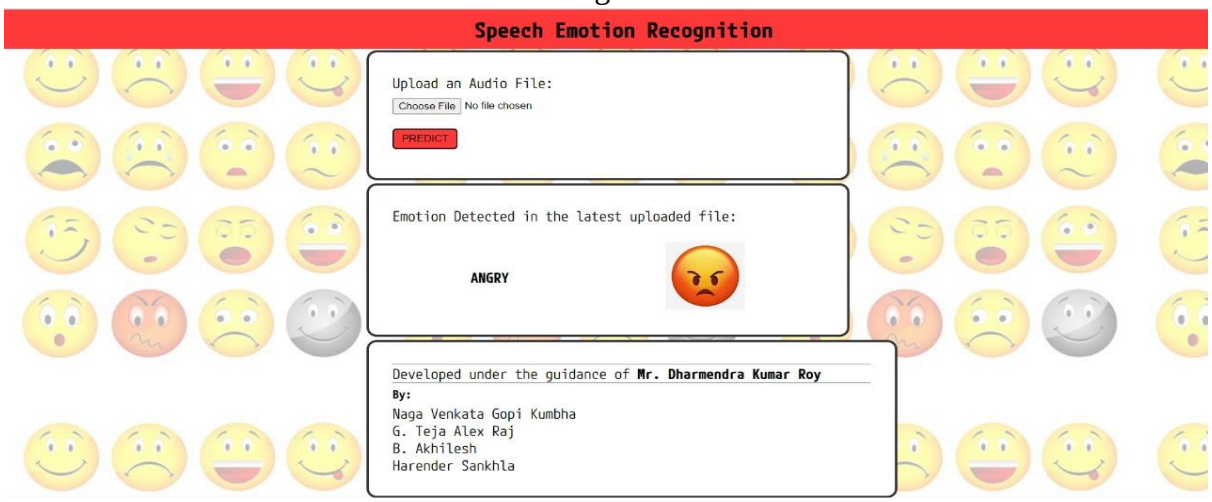

Fig. 8

International Journal of Engineering Technology and Management Sciences
Website: ijetms.in Issue: 3 Volume No.8May - June – 2024
DOI:10.46647/ijetms.2024.v08i03.020 ISSN: 2581-4621

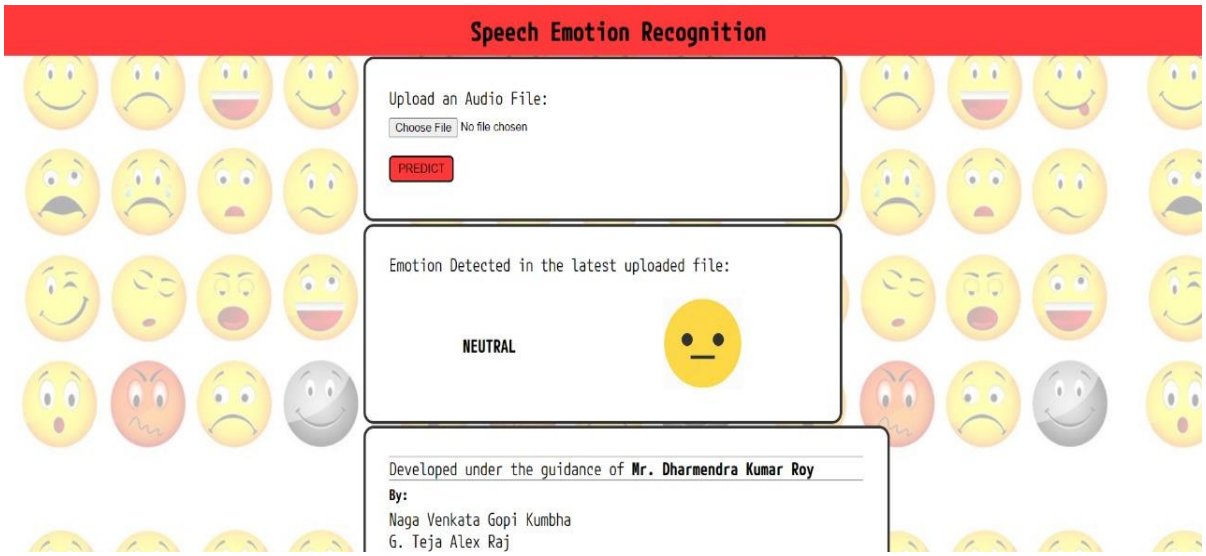

Fig. 9

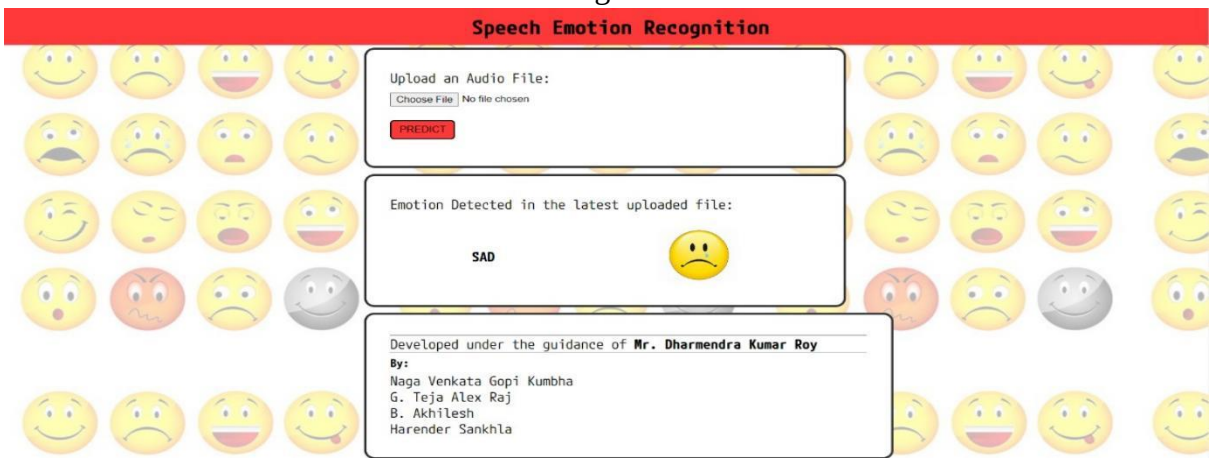

Fig. 10

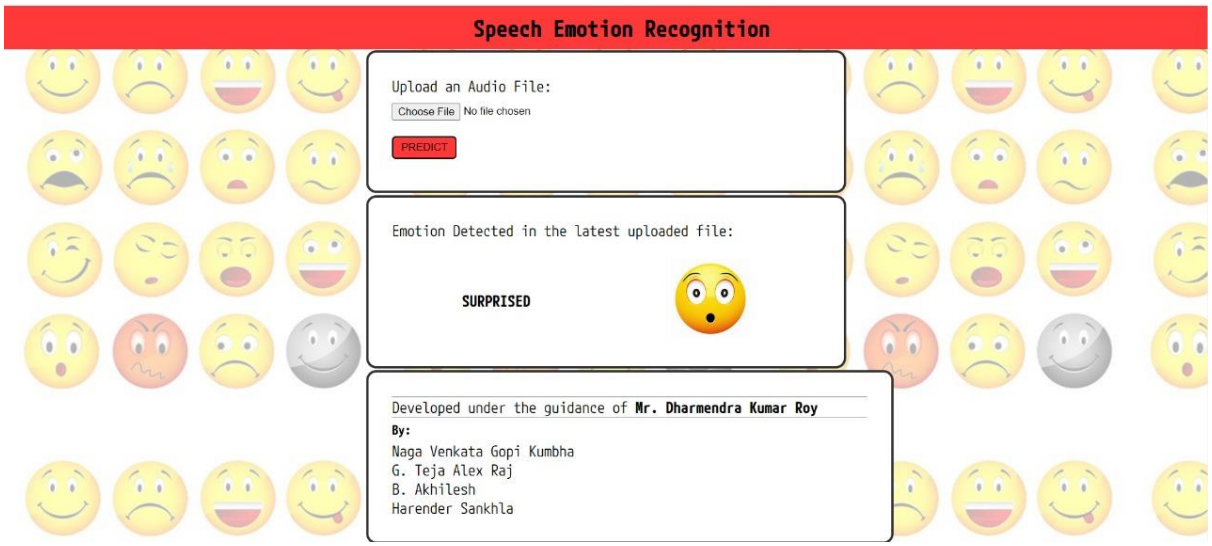

Fig. 11

International Journal of Engineering Technology and Management Sciences
Website: ijetms.in Issue: 3 Volume No.8May - June – 2024
DOI:10.46647/ijetms.2024.v08i03.020 ISSN: 2581-4621

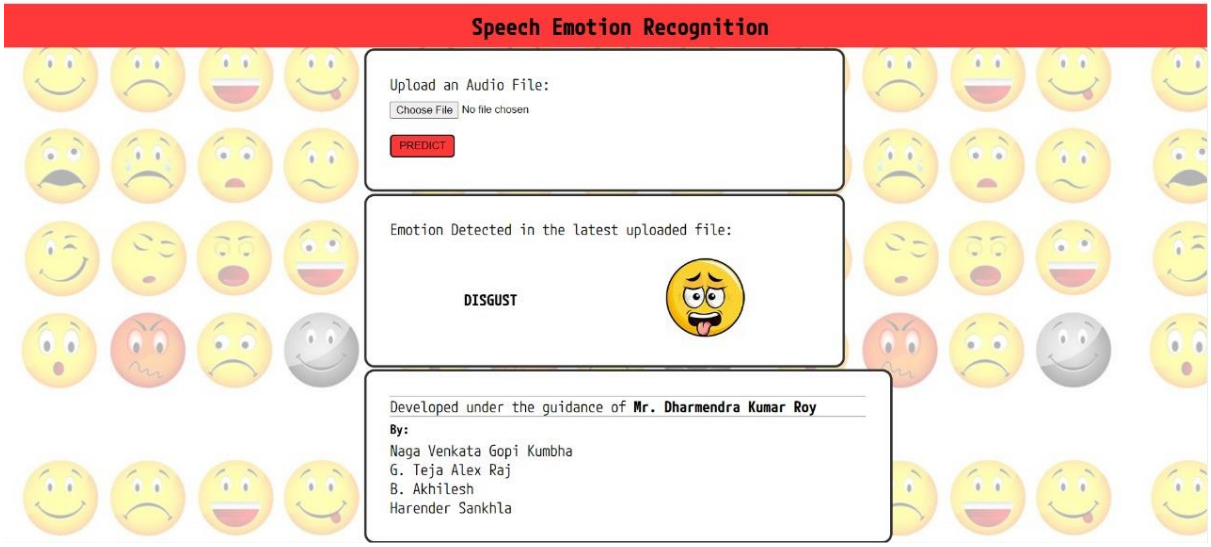

Fig. 12

References

[1] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In the International conference on machine learning, pages 214–223. PMLR, 2017.

[2] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzer-land, September 6-12, 2014, Proceedings, Part V 13, pages 740–755. Springer, 2014.

[3] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. International journal of computer vision, 123:32–73, 2017.

[4] Ruhul Amin Khalil, Edward Jones, Mohammad Inayatullah Babar, Tariqullah Jan, Mohammad Haseeb Zafar, Thamer Alhussain: Speech Emotion Recognition Using Deep Learning Techniques: A Review, IEEE Access (Volume: 7), 19 August 2019, DOC: 10.1109/ACCESS.2019.2936124

[5] M. Sakurai and T. Kosaka, "Emotion Recognition Combining Acoustic and Linguistic Features Based on Speech Recognition Results," 2021 IEEE 10th Global Conference on Consumer Electronics (GCCE), Kyoto, Japan, 2021, pp. 824-827, doi: 10.1109/GCCE53005.2021.9621810. keywords: {Emotionrecognition;Adaptation models;Conferences;Bit error rate;Speech recognition;Linguistics;Feature extraction;speech emotion recognition;linguistic features;acoustic features;speech recognition;language model adaptation},

[6] Maisy Wieman, Andy Sun: Analysing Vocal Patterns to Determine Emotion, 2017-09-22, Stanford University. Cs229. proj2014. Andy Sun, Maisy Wieman, Analysing Vocal Patterns to Determine Emotion.pdf

[7] M. S. Likitha, Sri Raksha R. Gupta, K. Hasitha, A. Upendra Raju: Speech based human emotion recognition using MFCC, 2017 International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET), 22 February 2018, DOC: 10.1109/WiSPNET.2017.8300161

[8] Jie Ju and Fang-Ai Liu. Multivariate time series data prediction based on att-lstm network. Applied Sciences, 11(20), 2021

[9] B. T. Atmaja and M. Akagi, "Speech Emotion Recognition Based on Speech Segment Using LSTM with Attention Model," 2019 IEEE International Conference on Signals and Systems

DOI:10.46647/ijetms.2024.v08i03.020 ISSN: 2581-4621

(ICSigSys), Bandung, Indonesia, 2019, pp. 40-44, doi: 10.1109/ICSIGSYS.2019.8811080. keywords: {Speech recognition;Emotion recognition;Feature extraction;Decoding;Computer architecture;Solid modelling;Conferences;voice segments;silence removal;speech emotion recognition;attention model},