SPEECH EMOTION RECOGNITION

Maniyar Mohammed Haroon
(1NC21CI028)
Dept. of CSE – AI&ML
Nagarjuna College of Engineering and Technology
Devanahalli, Bangalore
maniyarmohammedharoon@gmail.com

Mayank Singh (1NC21CI030)
Dept. of CSE – AI&ML
Nagarjuna College of Engineering and Technology
Devanahalli, Bangalore
mayanksingh2301@gmail.com

Riteesh M (1NC21CI040)
Dept. of CSE – AI&ML
Nagarjuna College of Engineering and Technology
Devanahalli, Bangalore
riteeshsaim19@gmail.com

Yammanuru Madhusudhana Reddy
(1NC21CI056)
Dept. of CSE – AI&ML
Nagarjuna College of Engineering and Technology
Devanahalli, Bangalore
yammanurumadhu654654@gmail.com

***Abstract*— Speech Emotion Recognition (SER) is an area of research related to the classification of emotions occurring in human speech. This is an area which contributes significantly toward applications in human-computer interaction, sentiment analysis, and monitoring mental health. This paper outlines a comprehensive approach to SER, based on techniques of machine learning. The system makes use of MFCCs, Chroma, and Mel Spectrograms, and a MLP classifier to classify emotions into the categories such as calm, happy, fearful, and disgust. The solution is a good robust preprocessing pipeline, real-time speech transcription, and visualization of performance metrics making it a useful tool for other domains. As such, the experimental results delineate the accuracy of the system and show an area of potential improvement towards classification of emotions.**

Keywords*— *Speech Emotion Recognition, Machine Learning, Feature Extraction, Multi-Layer Perceptron, Human-Computer Interaction.

I. INTRODUCTION

SER has recently gained much attention, especially with the emergence of today's systems that have been dealing with human-computer interaction. Emotions are an integral part of communication, and arming machines with emotion-recognizing capabilities could be very helpful in further enhancing the possibility of communicating machines to humans naturally. This paper deals with the development of a machine learning-based SER system in the context of audio signal processing with an emotion predicted from the input. Such technologies make fully possible the full deployment of SER systems in any application, be it virtual assistants and call center analytics, health diagnostics, or whatnot. More interestingly still, monitoring may be done as regards mental health or assisting one in therapy to improve better on educational tools having feedback on activity in terms of emotions. This requires processing and classifying advanced overlap in acoustic features due to emotion as well as scarcity in labeled datasets containing high-quality features and the presence of noise in backgrounds affecting SER systems, which makes the overall system impractical.

A. Problem Definition

Emotion recognition from speech poses a number of challenges. The most important challenge is the complexity of feature extraction that would represent emotional variability in speech. Variation in speech, depending on accents and speaking styles as well as contexts, complicates classification. The other important challenge is data imbalance, a biased training set, in which classes of emotion occur much less than others in the training data. It leads to poor generalization for those classes. Background noise and overlapping voices make the audio recording challenging to work with when trying to detect emotion. It also needs to work well on different accents, languages, and recording environments. The other challenge in the domain of emotion recognition is the necessity of real-time processing, where a balance of computational load is achieved for practical implementations. There is also the need for handling subtle expressions of emotions or mixed forms of emotions, which are hard to classify.

B. Objectives of this Project

The key objectives of this project are to design a machine learning-based system for emotion recognition from speech signals and to implement robust feature extraction techniques, such as MFCCs, Chroma, and Mel Spectrograms. Another objective is to train and evaluate an MLP classifier that can differentiate between multiple emotional states. Furthermore, the objective includes improving the model's generalization ability across different accents and recording conditions. This is achieved by making the system robust to changes in input quality and environmental noise. The other goal is to include speech transcription so that the spoken content will add another layer of insight into it, thereby making the system a multi-purpose analytical tool. The project also studies the effect of feature selection and optimization techniques on the overall accuracy of the system. The project also aims to develop real-time capabilities that can be practically applied while keeping the accuracy and efficiency of the system, thereby ensuring its applicability in dynamic real-world scenarios.

C. Limitations of the Project

Despite the encouraging outcome, there are some limitations for this proposed system. The recognition of four emotions only-calm, happy, fearful, and disgust-can be done with it. Extending to a complex or more subtle set of emotions is very hard. It strongly depends on the quality and diversity of the training dataset that influences generalization in real-world applications. It may degenerate in quality concerning its accuracy because of poor recording quality and external background noises due to the lack of advance mechanisms for noise reduction within it. There is no proper testing done with regard to its effectiveness in other languages and accents; thus, the challenge in its applicability in multilingual applications. This algorithm may also contain some limitations on the basis of real-time computing, which makes it infeasible for such low-resource devices. Feature extraction techniques must improve over time as it has scope for capturing better subtle emotional speech.

D. Organization of Documentation

This document is designed to provide a detailed and structured exploration of the research. The first segment introduces the concept of Speech Emotion Recognition, its significance, and the motivation behind this study. Following this, the problem definition elaborates on the specific challenges and gaps in the current state of the field. The objectives of the project are outlined next, highlighting the goals and intended outcomes of the research. The methodology section provides a comprehensive overview of the technical framework, including data preprocessing, feature extraction, and model training processes. The results and discussion section presents the experimental findings, supported by in-depth analysis and visualizations of performance metrics. Concluding the document is the summary of the study's contributions and recommendations for future research directions, emphasizing the potential advancements in SER systems.

E. Proposed Contribution

This paper enhances the domain of Speech Emotion Recognition using feature-rich extraction techniques such as MFCCs, Chroma, and Mel Spectrograms to augment the emotional signal representation. This paper also applies an adaptive learning rate Multi-Layer Perceptron in an effort to optimize prediction accuracy. The proposed system combines the emotions' predictions with speech transcription in multi-dimensional analysis, thereby showing a more detailed interpretation of the audio input. This model performance will also involve intuitive visualizations, such as confusion matrices and graphs for F1-scores for assessment and interpretation purposes. Its design, using this approach, will easily allow scalability in order to enhance it further with other languages, additional emotions, and realistic applications. With the address of the main feature extraction and generalization concerns in the preprocessing stage, the proposed work therefore guarantees an overall robust framework that paves the way for advancement in SER technology development. The proposed system, in turn, inherently depicts a scalable and practically feasible solution. This shall lead to the possibility of further human-computer interaction, sentiment analysis, and emotion-aware applications in fields like education, entertainment, and healthcare, among others.

II. Literature Survey

Ashish B. Ingale and D. S. Chaudhari (2012)
Speech Emotion Recognition (SER) systems aim to enhance human-machine interaction by identifying emotions from speech using features such as MFCC, LPCC, pitch, and energy. Classifiers like SVM, HMM, GMM, and ANN play a vital role in differentiating emotions such as anger, happiness, and sadness. However, cultural and linguistic diversity, variability in speaking styles, and transient emotions pose significant challenges. The study highlights the importance of using real-life emotional datasets to improve recognition accuracy. Applications include psychiatric diagnostics, intelligent systems, and driver safety enhancements.

Tiya Maria Joshy and Anjana S Chandran (2020)
Machine learning, especially deep learning models like CNN, RNN, and DBN, has revolutionized SER by automating feature extraction and handling large datasets effectively. Techniques like MFCC and pre-processing steps (e.g., noise removal) are critical in achieving reliable systems. Despite advancements, challenges such as subjectivity in emotional perception and complexities in data annotation persist. Integrating SER with applications like ASR has broadened its use in real-time human-computer interactions, including virtual assistants and sentiment analysis systems.

Kunal Bhapkar et al. (2021)
This study emphasizes robust feature extraction (e.g., MFCC and LPCC) and pre-processing techniques (e.g., silence removal, normalization) as foundational steps in SER. Classifiers like SVM and ANN are widely adopted, while hybrid models (e.g., DBN-SVM) have demonstrated improved accuracy. However, issues such as dataset variability, inconsistencies in emotional expression, and environmental noise remain challenges. Popular datasets like EMO-DB and RAVDESS are used for validating models, ensuring reproducibility. Applications of SER include real-time emotional monitoring in fields like healthcare,driver safety,customer-service.

Ö. Çağrı Dala (2023)
The study explores the role of advanced machine learning classifiers, including SVM, HMM, GMM, and ANN, in improving SER systems. Feature extraction focusing on prosodic and spectral attributes is critical, with datasets such as IEMOCAP and EMO-DB being widely used. The analysis suggests that combining classifiers, optimizing feature selection methods, and leveraging larger, diverse datasets significantly enhance accuracy. SVM is highlighted as the most effective classifier in several scenarios. The study also underlines the role of database quality, linguistic factors, and natural emotion variability in influencing recognition outcomes.

Babak Basharirad and Mohammadreza Moradhaseli (2017)
In their comprehensive literature review on speech emotion recognition, the authors analyzed various methods focusing on feature extraction, classification techniques, and dataset preparation. They evaluated the effectiveness of common datasets like Berlin and AIBO and highlighted challenges such as feature variability due to linguistic and cultural diversity. Techniques like MFCC, LPCC, and modulation spectral features were identified as foundational for emotion recognition. Classification methods such as support vector machines (SVMs), hidden Markov models (HMMs), and neural networks were explored, showing varying levels of accuracy. They emphasized hybrid approaches for improved recognition rates and suggested the integration of multiple classifiers for future advancements.

Kun Han, Dong Yu, and Ivan Tashev (2014)
This work introduced a hybrid model combining deep neural networks (DNNs) and extreme learning machines (ELMs) for speech emotion recognition. The authors utilized DNNs to extract high-level features from raw audio data, transforming segment-level outputs into utterance-level features. ELMs were used for classification, proving effective for smaller datasets. Their method achieved a 20% relative accuracy improvement over traditional methods like HMMs and SVMs. The study also demonstrated the advantage of DNNs in learning invariant features, reinforcing the role of deep learning in advancing emotion recognition.

Bagus Tris Atmaja and Masato Akagi (2019)

This study explored speech emotion recognition using a Long Short-Term Memory (LSTM) network combined with an attention mechanism. The authors emphasized the significance of voiced speech segments for emotion recognition, eliminating irrelevant silence through preprocessing. Their method integrated silence removal with attention-based LSTMs to focus on emotionally relevant parts of speech, achieving improved recognition accuracy on the IEMOCAP dataset. This approach highlighted the role of selective attention in enhancing deep learning-based emotion recognition systems.

Xubo Liu et al. (2020)

The authors proposed a speech emotion detection (SED) framework that identifies emotional classes and their temporal locations in continuous speech. They employed a sliding window approach for feature extraction and used Artificial Neural Networks (ANNs) to model the emotional state. This method demonstrated improved detection accuracy on the EMO-DB dataset by effectively combining interval-level and window-level emotional features. The study emphasized the importance of temporal modeling in continuous emotion detection.

Seyedmahdad Mirsamadi et al. (2017)

The research introduced a recurrent neural network (RNN) with a local attention mechanism for automatic speech emotion recognition. By focusing on emotionally salient speech segments, the proposed method overcame challenges of irrelevant silences and neutral speech regions. Experiments on the IEMOCAP corpus showed superior accuracy compared to traditional methods, underscoring the effectiveness of attention-based temporal aggregation in improving emotion recognition.

Jouni Pohjalainen and Paavo Alku (2013)

This paper focused on the automatic detection of anger in telephone speech using robust autoregressive modulation filtering. The authors combined autoregressive models and Gaussian Mixture Models (GMMs) to capture long-term temporal dynamics and enhance robustness under noisy conditions. By using the Berlin database, the system demonstrated high accuracy in distinguishing anger from other emotions, highlighting its application potential in real-world scenarios like call centers.

III. Proposed System

A. Overview of the System

The SER system designed here is based on the classification of emotions from audio speech signals using advanced machine learning techniques. The system is designed to work on processing audio data to come up with meaningful features, which are then analyzed by a Multi-Layer Perceptron MLP classifier to predict the emotional state. It also integrates a speech transcription mechanism to extract textual content from the audio. That makes it represent a multi-dimensional analysis for the input. The system is modular in design, hence scalable and adaptable, suitable for applications in human-computer interaction, sentiment analysis, and healthcare.

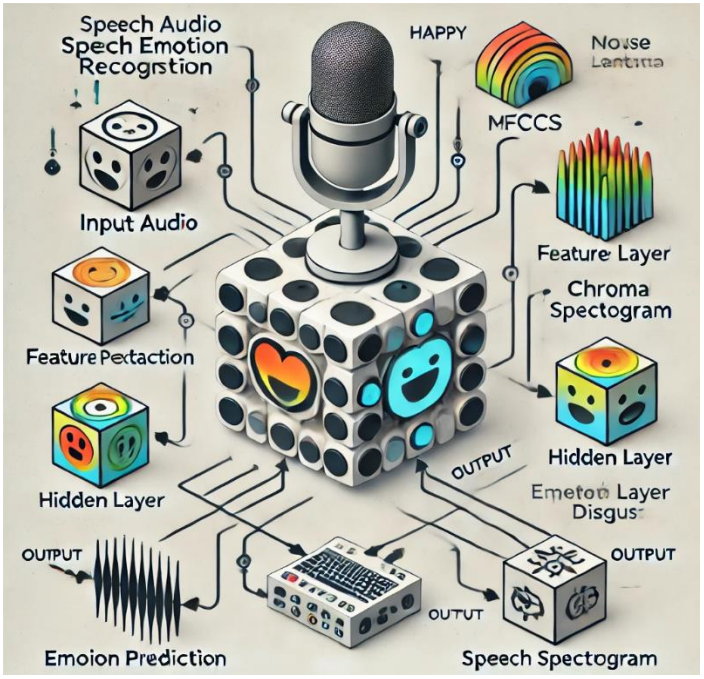
B. System Architecture

The architecture of the proposed system consists of the major components as listed below:

- Input Audio: The system accepts input audio files in .wav, .mp3, and .aac file formats.
- Preprocessing- It normalizes and enhances the input audio by applying noise reduction and converting files to.wav for uniformity.
- Feature Extraction: Critical features from the audio are extracted using techniques like Mel-Frequency Cepstral Coefficients (MFCCs), Chroma, and Mel Spectrograms.
- Emotion Recognition Model: A multi-layer perceptron (MLP) classifier is used to classify the extracted features into the corresponding emotion.

It shows the predicted emotion along with a transcription of the spoken content, which is very informative.

The following diagram visually represents the architecture of the proposed system:



"Figure 1: System Architecture of the Proposed Speech Emotion Recognition System."

C. Detailed Explanation of Each Component

- **Input Audio**

  The system will accept an audio recording from the user. Audio recordings can be given in any of a large set of file formats, from.mp3 to.aac. All input files are converted to the.wav format for ease of use with the audio processing libraries, as.wav is deemed the standard.

- **Preprocessing**

  This preprocessing step ensures quality in input data; therefore, this preprocessing stage may encompass:

  - Noise Reduction: This removes background noise to make the audio sound clearer.
  - Normalization It normalizes audio amplitudes to have an equal input.
  - Segmentation: Divide the long audio file into appropriate smaller segments with useful feature extraction.

- **Feature Extraction**

  Feature extraction is a critical stage in SER since the accuracy of the emotion classification relies heavily on the quality of the extracted features. The system utilizes three fundamental feature extraction techniques:

  - MFCCs (Mel-Frequency Cepstral Coefficients): This captures the spectral properties of speech, which are very vital in distinguishing the emotions.
  - Chroma Features: It is the harmonic and pitch content of an audio.
  - Mel Spectrograms: It gives a time-frequency representation, providing insights into the intensity and distribution of energy across different frequencies.

The extracted features are grouped together into a single feature vector, which will be used as input to the classifier.

- **Emotion Recognition Model**

  The system uses a Multi-Layer Perceptron (MLP) classifier to predict emotions. The MLP is configured with:

  - Input Layer: Accepts the concatenated feature vector.
  - Hidden Layer:It is comprised of three layers of 256, 128, and 64 neurons respectively and applies ReLU activation.
  - Output Layer: It gives the predictions for the observed emotions: calm, happy, fearful, and disgust.

This ensures that the model learns to make distinctions between these emotional states as it trains from the labeled dataset.

- **Speech Transcription**

  Apart from emotion prediction, the system transcribes the audio into text using a speech recognition API. This will add context to the emotion by analyzing the spoken words, which may be useful for applications in sentiment analysis and dialogue monitoring.

- **Output**

  The system produces an overall output that contains:

  - Predicted Emotion: It shows the emotion classified by the MLP model.
  - Speech Transcription: Displays the text content of the audio.
  - Performance Metrics: It gives visual insights such as confusion matrices and F1-score graphs to evaluate the performance of the model.

D. Advantages of the Proposed System

The proposed system has several advantages:

- Scalability: Modular design allows new emotions, languages, or applications to be added.
- Real-time Processing: Optimized for efficient real-time operation, making it practical for deployment in interactive systems.
- Multi-functionality: It integrates emotion recognition with speech transcription for a richer analysis of audio inputs.
- Robustness: Implements preprocessing steps to handle noise and variations in audio quality effectively.

E. WorkFlow Diagram

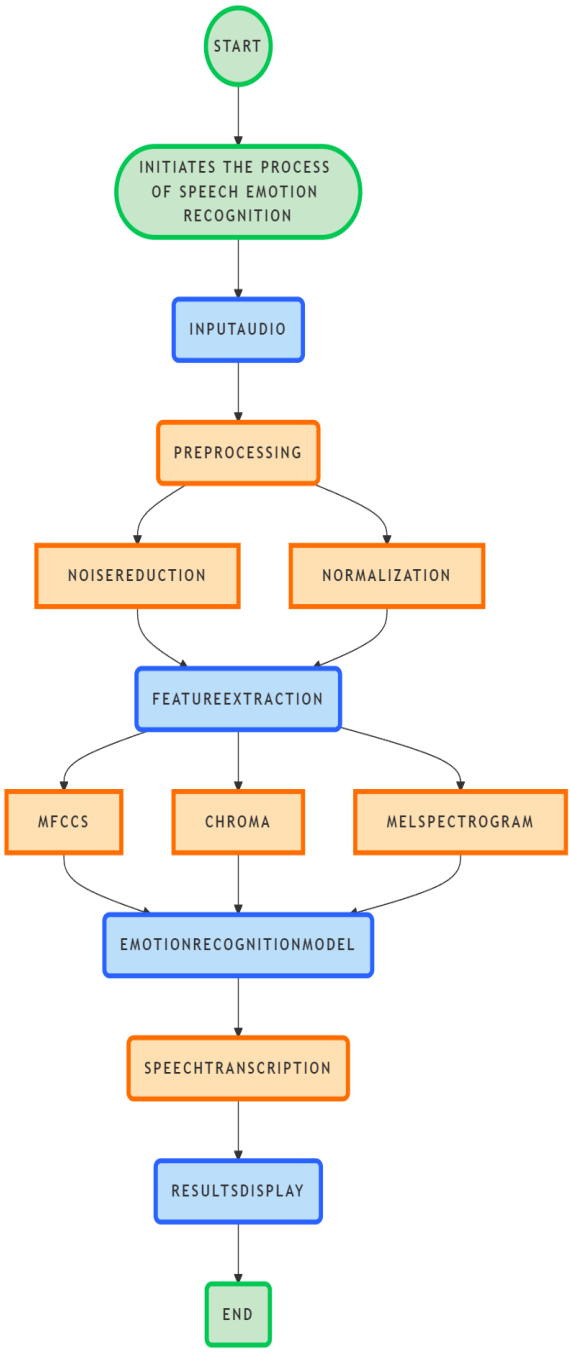


Figure 2 : System Workflow

IV. Results and Discussions

This section evaluates the performance of the proposed Speech Emotion Recognition (SER) system based on classification metrics, F1-score analysis, and frontend usability. The findings highlight the system's strengths, limitations, and areas for future improvements, with visual support from performance diagrams and the frontend interface-design.

A. *Overview of Result*

The proposed SER system achieved an overall accuracy of 74%, demonstrating its effectiveness in identifying emotions like "calm," "disgust," "fearful," and "happy." The capability of the system to process speech signals and extract meaningful features for classification provides a solid foundation for real-world applications in sentiment analysis and human-computer interaction.

Key Results:

The highest classification performance was observed for "calm," with an F1-score of 0.89, indicating distinct acoustic features for this emotion. While overlapping acoustic features characterize other emotions too, the happiness emotion was least performing with its F1 score at 0.61.

B. *Classification Metrics and F1-Score Analysis*

The **F1-score results (Figure 3)** and corresponding diagram (**Figure 4**) provide a detailed view of the system's classification performance for each emotion.

Classification Report:

	precision	recall	f1-score	support
calm	0.91	0.88	0.89	57
disgust	0.78	0.73	0.75	48
fearful	0.61	0.73	0.67	37
happy	0.62	0.60	0.61	50
accuracy			0.74	192
macro avg	0.73	0.73	0.73	192
weighted avg	0.75	0.74	0.74	192

Figure 3 : F1-Score Results

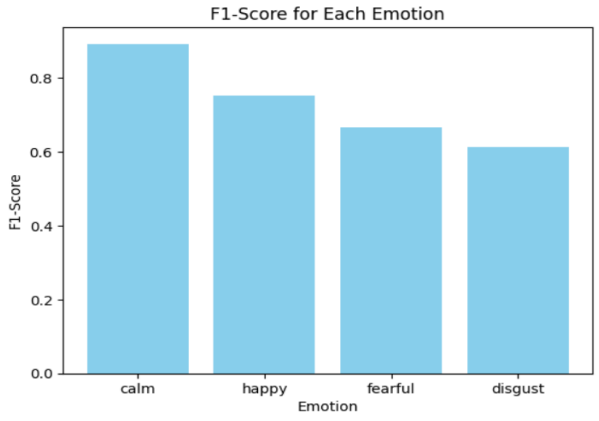


Figure 4 : F1 – Score Diagram

The diagram visually illustrates the system's F1-scores for different emotions, highlighting the model's strengths and areas requiring improvement. The "calm" class shows the most stable and high F1-score, while "happy" lags due to overlapping features with other emotions.

C. *Error Analysis Using Confusion Matrix*

The confusion matrix gives a blow-by-blow of how the system performs across actual and predicted emotion classes; Figure 5.

Key Observations:

- Minimal mis-classification for "calm," with few instances labeled as "happy" or "fearful."
- Notable confusion between "happy" and "calm," probably because of similarities in tonal expressions.
- Overlap between "fearful" and "disgust," reflecting common acoustic intensities and frequency ranges.

This analysis thus calls for refining the feature extraction methods to deal with overlapping feature spaces, which would improve the classification between challenging pairs like "happy" versus "calm."

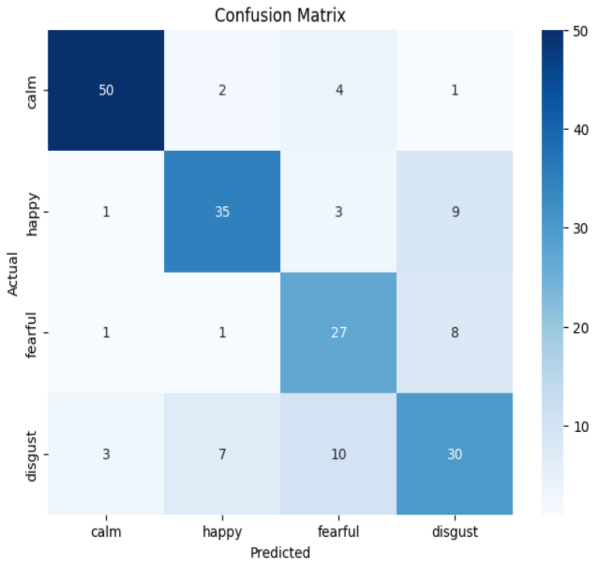


Figure 5: Confusion Matrix

D. *Frontend Interface Design*

Figure 6 depicts the friendly front-end interface through which a practical usability of the SER system can be explained. The design encompasses:

- Real-time speech recording module.
- Emotion prediction output is displayed dynamically.
- Visualization of performance metrics for better interpretability.

The interface is designed to be accessible and user-friendly, thus making the system feasible for applications such as virtual assistants and mental health diagnostics. The inclusion of a frontend system

bridges the gap between technical performance and end-user interaction, thus providing a complete solution.

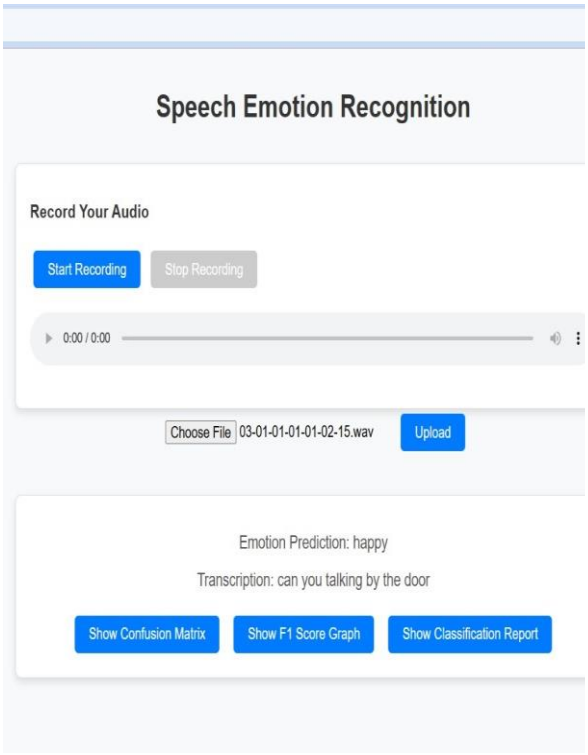


Figure 6: Frontend Interface

E. Challenges and Limitations

Although the results were very good, some issues were encountered.

- Overlapping Acoustic Features: The tonal features of "happy" and "calm" were similar, causing incorrect classifications.
- Data Imbalance: The distribution of samples between classes was imbalanced, resulting in the biased prediction of dominant classes.
- Noise Sensitivity: Background noise in the recordings poses a problem in classification, which may require more sophisticated noise removal techniques.
- Labeled Dataset: Diversity in accents and languages was sparse throughout the dataset and could affect model generalizability in real settings.

F. Future Directions

The findings open avenues for further research and improvements:

- Expansion of the Dataset: A larger, more diverse dataset with balanced samples for each emotion.
- Advanced Models Deep learning architectures such as CNNs and RNNs incorporate structures to give much more robust feature extraction and temporal analysis.
- Enhanced Noise Handling: Incorporate noise handling mechanisms to enhance performance under noise.
- Multilingual Support: Extending the system's capabilities to recognize emotions across multiple languages.

V. Conclusion

In this paper, we considered the state of the art in machine learning techniques for SER. We discuss the SER system to achieve high-precision emotion classification. The robust methods of feature extraction used in this system are MFCCs, Chroma, and Mel Spectrograms. It made use of a classifier known as Multi-Layer Perceptron (MLP) with substantial accuracy in distinguishing the emotions of speech signals. Experimental results yielded an accuracy of 74% for the classification, and "calm" showed to be the best performance here, but "happy" seems to have a lot of scope for improvement since it shares some acoustic features. Further analysis via the confusion matrix and F1-score revealed data imbalance and noise sensitivity.

Further development in the user-friendly front-end interface would further prove the applicability in reality, thereby making possible easy interaction as well as practicing usability. A few limitations regarding sensitivity with noise, an emotional class as limited, and troubles in multilingual environments were identified. These depict the need to improve the application in the following aspects: improvement of the available dataset, adoption of advanced deep techniques, and infusion of noise-reduction mechanisms into the system.

Some promise is exhibited in the findings of the research toward a functioning SER system inside applications such as human-computer interaction, sentiment analysis, or mental health diagnosis. This line of work creates a basis of further development towards the field. In the process of refining this developed system, accuracy, robustness, and generalizability should be boosted so that such a system might become more possible to be employed as a tool for emotion recognition in various application domains.

VI. References

[1] B. Tris Atmaja and M. Akagi, "Speech Emotion Recognition Based on Speech Segment Using LSTM with Attention Model," in Proceedings of the International Conference on Signals and Systems (ICSIGSYS), Nomi, Japan, July 2019 . The author presents an novel approach integrating the removal of silences along with an attention LSTM network in a way that leads to improvement over the existing literature in terms of accuracy specifically towards the IEMOCAP database.

[2] X. Liu, Y. Mou, Y. Ma, C. Liu, and Z. Dai, "Speech Emotion Detection Using Sliding Window Feature Extraction and ANN," in Proceedings of the International Conference on Signal and Image Processing (ICSIP), Beijing, China, March 2020. This paper presents a new framework that combines sliding window-based feature extraction with artificial neural networks for emotion detection in continuous speech, achieving the best performance on the EMO-DB dataset.

[3] S. Mirsamadi, E. Barsoum, and C. Zhang, "Automatic Speech Emotion Recognition Using Recurrent Neural Networks with Local Attention," in Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), New Orleans, LA, USA, March 2017.

This study highlights the importance of attention mechanisms in recurrent neural networks for focusing on emotionally salient speech regions, significantly enhancing recognition accuracy on the IEMOCAP dataset.

[4] J. Pohjalainen and P. Alku, "Automatic Detection of Anger in Telephone Speech with Robust Autoregressive Modulation Filtering," in Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Vancouver, Canada, May 2013. This article deals with the detection of anger in telephone speech using autoregressive modeling that improves robustness against noise. The results, based on experiments with the Berlin emotional speech database, showed a high level of reliability.

[5] K. Han, D. Yu, and I. Tashev, "Speech Emotion Recognition Using Deep Neural Network and Extreme Learning Machine," in Proceedings of the Interspeech Conference, Singapore, September 2014. This work shows that DNNs are used for feature extraction and ELMs for classification, with a 20% improvement in accuracy over state-of-the-art methods on benchmark datasets.

[6] B. Basharirad and M. Moradhaseli, "A Comprehensive Review of Speech Emotion Recognition Techniques,". This review comprehensively discusses the evolution of speech emotion recognition methods, highlighting challenges like feature variability and dataset diversity, and proposing hybrid classifier approaches for improved recognition performance.

[7] Ashish B. Ingale and D. S. Chaudhari, 2012.
"Speech Emotion Recognition," International Journal of Soft Computing and Engineering.
This paper covers the classification by SVM, HMM, and GMM using features such as MFCC and LPCC while highlighting cultural and linguistic-based variability cause and their application in psychiatric diagnosis and driver safety systems.

[8] Tiya Maria Joshy and Anjana S Chandran, 2020.
"Speech Emotion Recognition Literature Review," International Journal of Creative Research Thoughts, IJCRT,
This research illustrates how deep models such as CNN and RNN may be required in SER tasks for feature automatization. As an added highlight, the lack of annotation concerning emotional data further brings to point a better convergence of such mechanisms into real systems, including applications of virtual assistance and sentiment analysis programs.

[9] Kunal Bhapkar et al. (2021).
"Speech Emotion Recognition: A Survey," International Research Journal of Engineering and Technology (IRJET),
It discuses the techniques of feature extraction, like MFCC and the pre-processing steps, including noise removal to improve the accuracy of SER. Further, it reviews classifiers such as ANN and hybrid DBN-SVM models. Here, it's worth mentioning the importance of datasets such as EMO-DB and RAVDESS for the validation of models
.

[10] Ö. Çağrı Dala (2023).
"A Literature Review on Emotion Recognition in Speech,"
This article discusses the state of the art of machine learning classifiers like SVM, HMM, and GMM in recognizing emotions from the speech data on IEMOCAP and EMO-DB datasets. This paper suggests combining classifiers, optimising feature selection, and reducing challenges posed by linguistic and cultural differences to improve the SER system.