WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

SPEECH EMOTION RECOGNITION USING MACHINE LEARNING

Prof. A. V. Pande [1]***,Yash Ganjare*** [2]***, Priyanshu Bodele*** [3]***, Abhishek Patane*** [4]***, Vivek Shete*** [5]***, Parth Abruk*** [6]***.***

[1]Assistant Professor,
[2,3,4,5,6] UG Students,
Department of Computer Science and Engineering,
Sipna College of Engineering and Technology, Amravati, Maharashtra, India

ABSTRACT –

Speech Emotion Recognition (SER) system using machine learning techniques. It utilizes the RAVDESS dataset to identify emotions such as happiness, sadness, anger, and calm from speech. Key features like MFCC, Chroma, and Mel spectrogram are extracted to represent the emotional content of audio signals. A Multilayer Perceptron (MLP) classifier is trained to classify these features into predefined emotion categories. The system is developed using Python and integrated with a Flask-based web interface. It achieves an impressive accuracy of 82% in emotion detection. This model has applications in virtual assistants, healthcare, and enhancing human-computer interaction through empathetic responses.

Feature extraction is efficiently handled using the Librosa library, streamlining the processing pipeline. Noise reduction techniques are also applied to improve recognition accuracy in real- world environments. The system's UI allows users to input speech and get real-time emotion feedback. Overall, this project contributes to the advancement of empathetic AI technologies.

Keywords - Speech Emotion Recognition, Machine Learning, MFCC, MLP Classifier, RAVDESS Dataset, Audio Feature Extraction, Chroma, Mel Spectrogram, Human- Computer Interaction, Emotion Detection, Python, Flask.

Introduction

Speech Emotion Recognition (SER) using machine learning to classify emotions from human speech. Emotions such as happiness, sadness, anger, fear, calm, and others are detected using audio signals. The system is built using the RAVDESS dataset, which contains labeled emotional speech samples from male and female actors. Key audio features like Mel Frequency Cepstral Coefficients (MFCC), Chroma, and Mel Spectrogram are extracte using the Librosa library.

These features help capture pitch, tone, and frequency patterns essential for identifying emotions. A Multilayer Perceptron (MLP) classifier is used to train the model for emotion classification. The model achieves an accuracy of 82%, showcasing its effectiveness. The system includes data preprocessing, noise reduction, and feature extraction to enhance model performance. The user interface is developed using Flask, allowing users to input speech and receive real-time emotion feedback. Visualizations like waveforms and spectrograms are also integrated. The backend processes audio through various transformations to isolate emotional patterns

This SER model offers real-time emotion detection capabilities. It has potential applications in human-computer interaction, mental health monitoring, virtual assistants, and education. The project demonstrates how machines can understandhuman emotions through voice. It contributes to building empathetic, intelligent systems. Future enhancements can include multimodal emotion detectionusing facial and textual data.

Literature Survey

In recent years, various systems and technologies have been developed to reduce food waste and manage surplus food efficiently. These research contributions highlight different approaches, including web platforms, IoT systems, policy frameworks, deep learning, mobile apps, and gamified user engagement.

Utkarsh Garg et al. (2020) proposed a method combining Mel Frequency Cepstral Coefficients (MFCC), Mel Spectrogram, and Chroma features to improve emotion classification. Their feature fusion approach significantly enhanced the recognition accuracy in SER systems.

Anurish Gangrade et al. (2022) introduced a Deep Belief Network (DBN) integrated with a Convolutional Neural Network (CNN) model for emotion detection. This deep learning-based architecture achieved high recognition rates by capturing complex emotional patterns in speech.

Girija Deshmukh et al. (2019) emphasized using MFCCs and energy features with the RAVDESS dataset. Their research demonstrated effective prediction of emotional states through detailed feature extraction and machine learning classification.

S.G. Shaila et al. (2023) utilized models such as Random Forest and Multilayer Perceptron (MLP) for emotion classification. Their study highlighted the effectiveness of classical machine learning algorithms in processing emotional speech data.

T. Kishore Kumar (2017) proposed a novel fusion of Teager Energy Operator (TEO) and MFCC called Teager-MFCC (T-MFCC) to improve recognition of stressed emotions, showing superior results compared to standard MFCC-based methods.

Chen Caihua (2019) focused on multi-modal Mandarin SER using Support Vector Machines (SVM). The study addressed speech signal pre-processing, feature fusion, and classification for enhanced emotional detection accuracy.

Husbaan Attar et al. (2022) developed a real-time SER system capable of analyzing continuous speech, with applications in online education and personalized feedback. Their approach improved engagement and interaction in learning platforms.

Vinita Chugh et al. (2021) examined the challenges of emotion differentiation and the limitations of current speech datasets. They stressed the need for richer datasets and advanced feature sets to improve model robustness.

Analysis Of Problem

The increasing reliance on intelligent systems has highlighted a significant limitation— machines often fail to understand the emotional context of human speech. Traditional speech processing systems focus solely on linguistic content, overlooking the paralinguistic elements such as tone, pitch, and rhythm that convey emotional information. This lack of emotional awareness restricts the effectiveness of applications like virtual assistants, customer service bots, and mental health monitoring systems.

Human speech contains rich emotional cues that are vital for natural and empathetic communication. However, capturing and interpreting these emotions is challenging due to the variability in speech patterns across individuals, languages, and cultural backgrounds.

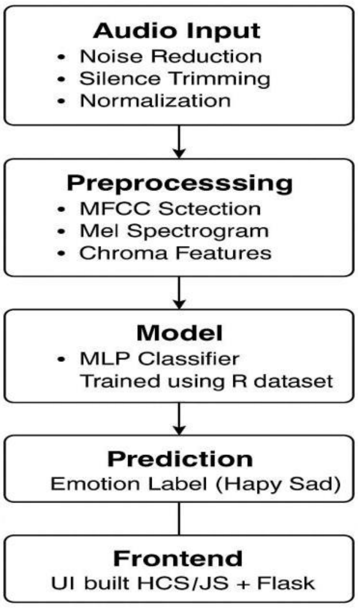
Emotions are subjective and often expressed differently by different people, making classification a non-trivial task.

Another key problem is the absence of real-time, accurate, and scalable emotion detection systems. Many existing models either lack precision or are computationally intensive, making them unsuitable for real-world applications. Furthermore, background noise, poor audio quality, and overlapping speech can affect the accuracy of emotion recognition models.

To address these challenges, there is a need for a robust system that can accurately identify emotions from speech using efficient machine learning algorithms and reliable feature extraction techniques. By implementing such a system, we can bridge the emotional gap in human-computer interaction, leading to more responsive and human-centric technology.

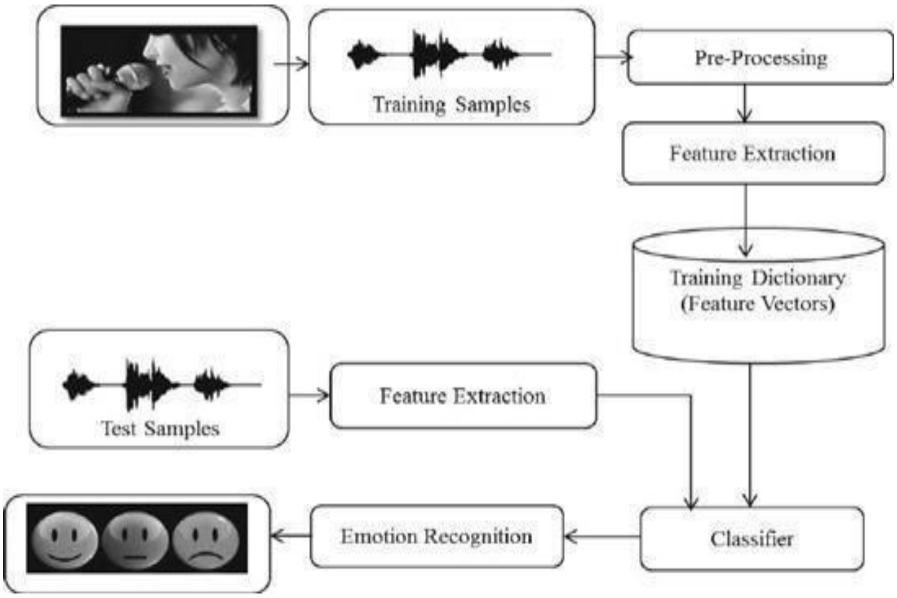
System Design

The Speech Emotion Recognition (SER) system is designed to detect and classify human emotions from spoken audio using machine learning techniques. The process begins with the user providing input through a microphone or by uploading a `.wav` audio file. Once the audio is captured, it undergoes a preprocessing phase where background noise is reduced, silence is trimmed, and the audio signal is normalized to ensure clarity and uniformity. After cleaning, the system extracts meaningful features from the speech using methods such as Mel- Frequency Cepstral Coefficients (MFCC), Mel Spectrogram, and Chroma features. These features capture the tonal, spectral, and pitch-related properties of the speech signal, which are critical for emotion analysis



The extracted features are then passed to a trained machine learning model, specifically a Multi-Layer Perceptron (MLP) classifier. This model has been trained using the RAVDESS dataset, which includes audio samples labeled with emotions such as happiness, sadness, anger, fear, calm, and more. The classifier processes the input and predicts the emotion conveyed in the speech. This output is then displayed to the user through a web interface built using HTML, CSS, JavaScript, and Flask. The interface provides options to record or upload speech, visualize waveform and spectrogram data, and view the predicted emotion along with a confidence score. Overall, this system integrates audio processing, feature extraction, machine learning, and web development to enable real-time emotion detection from speech.

System FlowChart



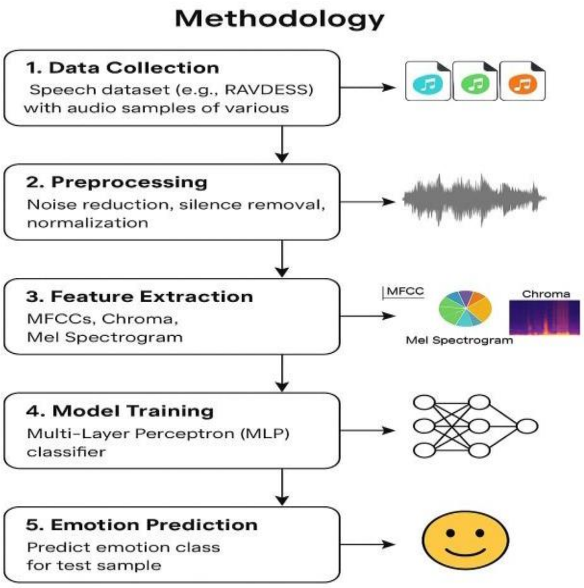
The flowchart visually represents the overall architecture of a **Speech Emotion Recognition (SER)** system. It begins with the collection of **training samples**, which are audio recordings of human speech. These recordings undergo **pre-processing** to clean the audio by reducing noise and normalizing signals. Following this, **feature extraction** techniques such as MFCCs are applied to transform raw audio into useful numerical data.

These extracted features are stored in a **training dictionary**, forming the basis for training the emotion classifier. On the other side, **test samples** also go through feature extraction. The features from the test samples are then passed to the **classifier**, which compares them with the training dictionary to recognize patterns.

Based on this, the classifier performs **emotion recognition**, identifying emotions like happy, sad, or angry. The results are finally displayed in a **visual format** representing the detected emotion. This flowchart outlines the key stages from raw speech to emotion output in a machine learning-driven system

Methodology

The methodology for Speech Emotion Recognition (SER) involves several systematic stages, from audio input to emotion classification. This pipeline ensures that raw speech signals are effectively processed and analyzed for emotion detection using machine learning models.


Methodology
1. Data Collection
Speech dataset (e.g., RAVDESS) with audio samples of various
2. Preprocessing
Noise reduction, silence removal, normalization
3. Feature Extraction
MFCCs, Chroma, Mel Spectrogram
MFCC
Chroma
Mel Spectrogram
4. Model Training
Multi-Layer Perceptron (MLP) classifier
5. Emotion Prediction
Predict emotion class for test sample

1. **Data Collection**

The system uses a speech dataset such as **RAVDESS**, which contains professionally recorded audio samples representing various emotions including happy, sad, angry, calm, fearful, and more. These are divided into **training** and **test** datasets.

2. **Preprocessing**

The audio signals are preprocessed to enhance quality and remove unwanted artifacts:

- **Noise Reduction** using filters.
- **Silence Removal** for trimming non-speech parts.
- **Normalization** to unify volume levels.

3. **Feature Extraction**

Speech signals are converted into meaningful numerical representations. The commonly used features include:

- **MFCC (Mel-Frequency Cepstral Coefficients)**
- **Chroma Features**
- **Mel Spectrogram**

These features capture pitch, tone, and frequency-related information essential for emotion recognition.

4. **Model Training**

Extracted features from the training samples are fed into a **machine learning model**, such as a **Multi-Layer Perceptron (MLP)** classifier. The model learns to identify patterns associated with different emotions based on feature vectors.

5. **Emotion Prediction**

When a test audio sample is submitted, it undergoes the same preprocessing and feature extraction steps. These features are passed to the trained model, which then predicts the emotion class. The output is displayed to the user with a confidence score**.**

Result

The Speech Emotion Recognition (SER) system developed in this project demonstrates the effectiveness of machine learning in identifying human emotions through speech. By utilizing the **RAVDESS dataset**, the system is capable of classifying emotions such as **happiness, sadness, anger, calm, and fear** based on audio input.

Key findings from the system's implementation are as follows:

- **Accuracy Achieved**: The system achieved a notable **accuracy of 82%** using the **Multilayer Perceptron (MLP) classifier**, trained on extracted features like **MFCC**, **Chroma**, and **Mel Spectrogram**.
- **Feature Extraction**: Employing the **Librosa library** allowed for efficient and robust feature extraction. The combination of MFCC, Chroma, and Mel spectrogram proved effective in capturing critical tonal and pitch variations necessary for accurate emotion classification.
- **Noise Reduction Impact**: Preprocessing techniques such as **noise filtering**, **silence trimming**, and **signal normalization** enhanced the model's performance by improving feature clarity, especially in real-world audio samples.
- **Real-Time Feedback**: Integration with a **Flask-based web interface** enabled real-time audio input and emotion detection. The interface not only provides users with visualizations like **waveforms and spectrograms**, but also returns the **predicted emotion** and its **confidence score** instantly.
- **System Usability**: The system supports **both live recording** and **file upload**, making it user-friendly and suitable for diverse environments, including **virtual assistants**, **customer service bots**, **mental health applications**, and **interactive education platforms**.
- **Dataset Performance**: Using the **RAVDESS dataset**, which includes well-annotated emotional speech from multiple speakers, ensured that the model was exposed to a variety of emotional tones, leading to generalized performance across different voices and genders.

Conclusion

The project on Speech Emotion Recognition (SER) using machine learning represents a significant endeavour at the intersection of artificial intelligence and human psychology. Emotion, a cornerstone of human interaction, is intricately woven into speech, making its recognition crucial for effective communication. By harnessing machine learning algorithms and computational methods, we aim to bridge the gap between human emotions and technological interfaces. This project addresses the growing need for empathetic computing, where machines can not only understand but also respond to human emotions in real-time. The motivation behind Speech Emotion Recognition (SER) using machine learning offers numerous real-life applications, including human-computer interaction, customer sentiment analysis, and mental health assessment. SER enables virtual assistants to adapt responses based on user emotions, enhances customer service by analysing emotional content in interactions, and aids in early detection of mood disorders. Successful implementations include real-time analysis in call centres, personalized learning in education, and proactive intervention in mental health care. These applications highlight SER's potential to revolutionize various domains, creating more empathetic and responsive technologies. By developing robust SER systems, we pave the way for more empathetic and responsive technologies that resonate with users on a deeper emotional level.

REFERENCE

1. Utkarsh Garg, Sachin Agarwal, Shubham Gupta, Ravi Dutt and Dinesh Singh, "Prediction of Emotions from the Audio Speech Signals using MFCC, MEL and Chroma," 12th International Conference on Computational Intelligence and Communication Networks, Nov 2020.
2. Anurish Gangrade, Shalini Singhal, "A Research of Speech Emotion Recognition Based on CNN Network," SKIT Research Journal, VOLUME 12, ISSUE 1, July 2022.
3. Girija Deshmukh, Apurva Gaonkar, Gauri Golwalkar, Sukanya Kulkarni, "Speech based Emotion Recognition using Machine Learning," 3rd International Conference on Computing Methodologies and Communication (ICCMC), March 2019.
4. S. G. Shaila, A. Sindhu, L. Monish, D. Shivamma, and B. Vaishali, "Speech Emotion Recognition Using Machine Learning Approach," ICAMIDA 2022, ACSR 105, pp. 592– 599, May 2023
5. T. Kishore Kumar, "Stressed Speech Emotion Recognition using feature fusion of Teager Energy Operator and MFCC," IEEE – 40222 8th ICCCNT, July 2017.
6. Chen Caihua, "Research on Multi-modal Mandarin Speech Emotion Recognition Based on SVM," 2019 IEEE International Conference on Power, Intelligent Computing and Systems (ICPICS), July 2019.
7. Husbaan I. Attar, Nilesh K. Kadole, Omkar G. Karanjekar, Devang R. Nagarkar, Prof. Sujeet and More, "Speech Emotion Recognition System Using Machine Learning," InternationalJournal of Research Publication and Reviews, Vol 3, no 5, pp 2869-2880, May 2022.
8. Sonali T. Saste, Prof. S. M. Jagdale, "Emotion Recognition from Speech Using MFCC and DWT for Security System," International Conference on Electronics, Communication and Aerospace Technology ICECA, April 2017.
9. Ryota Sato, Ryohei Sasaki, Norisato Suga, Toshihiro Furukawa, "Creation and Analysis of Emotional Speech Database for Multiple Emotions Recognition," Proceedings of the 23rd Conference of the Oriental COCOSDA, Yangon, Myanmar, Nov 2020
10. Vinita Chugh, Shivanghee Kaw, Surabhi Soni, Varsha Sablani & Rupali Hande,
    "Speech Emotion Recognition System Using MLP," 2021 JETIR October 2021, Volume 8, Issue 10 www.jetir.org (ISSN-2349-5162), Oct 2021.
11. Kottilingam. Kottursamy, "A review on finding efficient approach to detect customer emotion analysis using deep learning analysis," Journal of Trends in Computer Science and Smart Technology, vol. 3, no. 2, pp. 95-113, July 2021.
12. Amrita Thakur, Pujan Budhathoki, Sarmila Upreti, Shirish Shrestha and Subarna Shakya, "Real Time Sign Language Recognition and Speech Generation," Journal of Innovative Image Processing, vol. 2, no. 2, pp. 65-76, June 2020..