IJNRD.ORG ISSN : 2456-4184
INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD) | IJNRD.ORG
An International Open Access, Peer-reviewed, Refereed Journal

SPEECH EMOTION DETECTION SYSTEM USING PYTHON

[1]Keetha Ajay Kumar, [2]Jalli Srujan Kumar, [3]Gotte Pavani, [4]Kalva Sai Abhinav, [5]Elijah Francis

[1]Student, [2]Student, [3]Student [4]Student, [5]Assistant Professor
[1]Computer Science and Engineering,
[1]Joginpally B R Engineering College, Hyderabad, India

Abstract: In recent years, there has been a growing interest in developing systems that can detect human emotions from speech. The Speech Emotion Detection System (SEDS) using Python is designed to analyze audio recordings and accurately classify the emotional state of the speaker. This project leverages machine learning and signal processing techniques to achieve its objectives, such as happiness, sadness, anger, fear, surprise, and neutrality.

I. INTRODUCTION

With advancing artificial intelligence, there is an element that makes human emotion the cornerstone of improving human-computer interaction. SER exploits the richness of human speech to recognize emotions, bringing about a new method to bridge technology and human behavior. The Python based **SERS** system is about analyzing audio recordings in order to classify emotions such as happiness, sadness, anger, fear, surprise and neutrality. The system uses advanced techniques from both machine learning and signal processing in order to decode an emotional tone that is embodied in speech. This application of the system is utilized in customer service, in mental health monitoring, as well as in IVR systems.

The methodology of this project includes extracting acoustic features like pitch, tone, speed, and energy from raw audio data, followed by machine learning algorithms to classify these features into emotional states. The system's capability to process different speech formats and adapt to real-time scenarios ensures its versatility and practicality. By dealing with the challenges of real-world applications such as noise sensitivity and scarcity of data, this project looks forward to overcoming the limitations of the current models. It shows the potential for emotion-aware AI systems, fostering innovation in domains where understanding human emotions is vital.

II. NEED OF THE STUDY.

Emotions form a principal component of human interaction, bearing paramount influence on the dynamics of social discourse, decision-making, and, in general, humans' well-being. Intelligibility of emotions in speech became highly valued recently due to its relevance to success with the concept of artificial intelligence (AI) in enhancing human-computer interaction. However, current systems are facing significant limitations in detecting emotions feasibly, especially in real scenarios. The **Speech Emotion Recognition System (SERS)** using Python overcomes these challenges by employing more advanced machine learning techniques as well as audio signal processing.
This paper will develop a strong and scalable solution that will help overcome the limitations of the current models, focusing on real-time emotion detection with practical applications. The research is contributing to the development of technologies that not only work efficiently but also align with human-centric design principles.

2.1 Data and Sources of Data

For the Speech Emotion Detection System Using Python, different datasets are used to train and evaluate emotion classification models. Among the data sources, one of the most prominent ones is the RAVDESS (Ryerson Audio-Visual Database of Emotional Speech and Song), which contains audio and video recordings of actors acting out various emotions, including happiness, sadness, anger, fear, and surprise. This dataset comprises a broad range of expressions, both neutral and extreme, making it ideal for training robust emotion recognition models.

3.3 Theoretical framework

The Speech Emotion Detection System Using Python is a creation of those principles; speech signal processing, machine learning, and emotion recognition. This model combines those disciplines in creating the process for classifying emotional state from an audio signal where techniques applied are drawn mainly from signal processing to develop features while the model based on the machine learning makes the decision about emotions.

III. RESEARCH METHODOLOGY

The Speech Emotion Detection System Using Python outlines the systematic process used to develop, implement, and evaluate the emotion detection system. It follows a structured approach that combines data collection, preprocessing, feature extraction, model development, and evaluation. Below is a detailed description of the methodology:

3.1 Data Collection

The first step in the research methodology is the collection of data. The system relies on publicly available speech emotion datasets, such as RAVDESS, EmoDB, CREMA-D, TESS, SAVEE, and others. These datasets contain labeled audio recordings of human speech, each annotated with emotional labels such as happiness, sadness, anger, fear, surprise, and neutrality. The collected datasets provide diverse speech samples, including various emotional expressions, languages, and speaking styles. These datasets serve as the foundation for training and testing the emotion detection system.

3.2 Data Preprocessing

Once the data is collected, the next step is preprocessing. Raw audio signals are typically noisy and may contain irrelevant background sounds, so preprocessing is essential to clean the data. The preprocessing phase includes several tasks:

- Noise Reduction: Using filters to remove background noise and irrelevant sounds from the speech signal, ensuring that only the important acoustic features are preserved.
- Segmentation: Breaking down the audio recordings into smaller segments, such as individual sentences or words, to facilitate better analysis and reduce complexity.
- Normalization: Ensuring that the audio data has a consistent volume level and sample rate, making it suitable for feature extraction and model training.

The goal of preprocessing is to prepare the raw speech data so that it is in a suitable format for extracting meaningful features while minimizing distortion caused by noise and inconsistencies.

3.3 Feature Extraction

Feature extraction is a crucial step in the methodology, as it converts raw audio signals into numerical representations that capture the relevant characteristics of speech. Various acoustic features are extracted from the preprocessed audio data:

- Mel Frequency Cepstral Coefficients (MFCCs): These coefficients represent the shape of the speech spectrum and are widely used in speech processing tasks because they mimic human auditory perception.
- Pitch: The fundamental frequency of the speech signal, which reflects the tone of the speaker and is highly correlated with emotional states.
- Energy: The intensity or loudness of the speech signal, which can vary significantly depending on the emotional state of the speaker.
- Spectral Features: Characteristics of the frequency spectrum that describe the voice's tonal quality and can provide insight into the emotional content of speech.

The extracted features are typically organized into feature vectors, with each vector representing a segment of the audio signal. These feature vectors are used as input for the machine learning model.

3.4 Model Development

The core of the system is the emotion classification model, which uses the extracted features to identify the emotional state of the speaker. The proposed system utilizes an MLP (Multilayer Perceptron) model, which is a type of artificial neural network. The model consists of multiple layers:

- Input Layer: The feature vector extracted from the speech signal is passed to the input layer of the MLP model.
- Hidden Layers: The model contains one or more hidden layers with ReLU activation functions, which allow it to learn complex relationships between the input features and the emotional states.
- Output Layer: The output layer is a softmax layer that outputs a probability distribution across the different emotion classes, such as happiness, sadness, anger, fear, surprise, and neutrality.

The MLP model is trained using labeled data (audio features with known emotion labels) through a process called supervised learning. During training, the model adjusts its weights to minimize the error between the predicted and actual emotional labels. The model is trained using a backpropagation algorithm, which iteratively updates the model's parameters to improve classification accuracy.

3.5 Model Evaluation

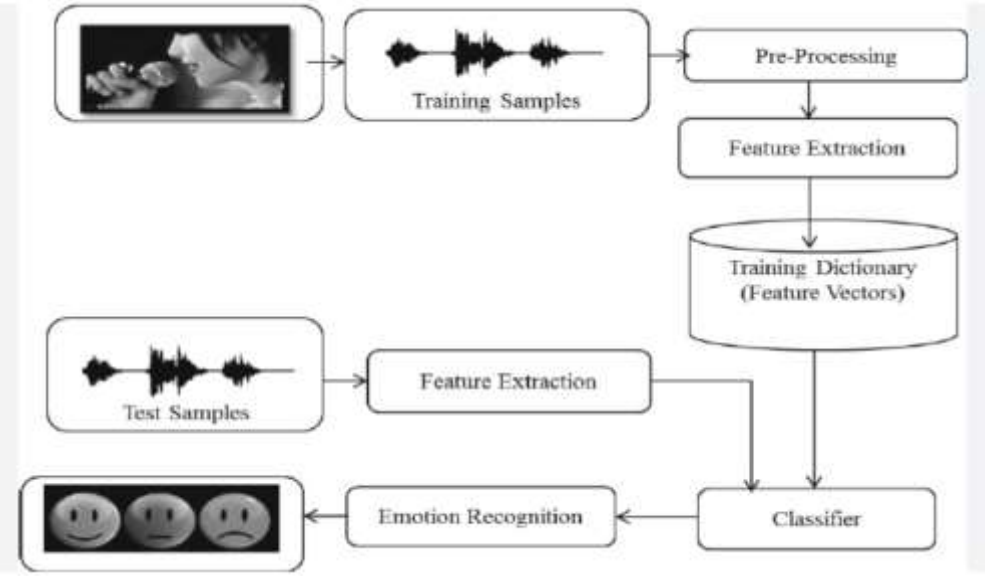
After training the model, it is evaluated using a separate test dataset that was not seen during the training phase. The evaluation phase assesses the performance of the model in terms of its ability to accurately classify emotions from unseen audio data. Common evaluation metrics used in this process include:

- Accuracy: The percentage of correctly classified instances out of the total instances.
- Precision, Recall, and F1-**Score**: These metrics provide deeper insights into the model's performance, especially in terms of handling imbalanced datasets or specific emotion classes.
- Confusion Matrix: This matrix provides a detailed breakdown of the model's performance for each emotion class, showing how often each emotion is correctly or incorrectly predicted.

Cross-validation techniques, such as k-fold cross-validation, can also be applied to ensure that the model's performance is robust and not biased by a specific training or test set.

IV. SYSTEM DESIGN

4.1 ARCHITECTURE:



4.1.1 Input Layer:

The system accepts raw audio files from the user or from pre-loaded datasets.

4.1.2 Preprocessing Layer:

The preprocessing layer is a crucial step in the Speech Emotion Detection System, where raw audio data is transformed into a format suitable for feature extraction and classification. The primary goal of this layer is to clean the data, standardize it, and prepare it for the machine learning models. The following steps are involved in this layer

4.1.3 Feature Extraction:

Feature extraction is a critical step in the Speech Emotion Detection System as it involves identifying and extracting meaningful attributes from the preprocessed audio data. These attributes, or features, help the machine learning models (like CNN, SVM, and Random Forest) distinguish between different emotions in the speech. In this system, we primarily use two types of feature extraction methods: traditional audio features (like MFCCs and Chroma) and deep learning-based features (using CNN).

- **CNN**: The deep learning model extracts features from the graphical representations, such as patterns in frequency and time.
- **MFCCs**: In parallel, MFCCs are extracted as additional features.

4.1.4 Classification Module:

The Classification Module is the core of the Speech Emotion Detection System, where the extracted features from the audio data are analyzed and classified into different emotional categories (e.g., "sad," "angry," "happy"). In this system, a combination of traditional machine learning models (SVM, Random Forest) and deep learning models (CNN) are used to ensure accurate emotion detection.

- **CNN**: A convolutional neural network classifies the emotion from the extracted features.
- **SVM and Random Forest**: These models are used as additional classifiers to cross-check the CNN's predictions.

4.1.5 Output Layer:

The **Output Layer** in the Speech Emotion Detection System is the final step where the classification results are generated after the model processes the input audio data. This layer is responsible for predicting the emotional state based on the features extracted by the CNN (or other classifiers like SVM and Random Forest). The output is a categorical classification of emotions such as "happy," "sad," "angry," or "neutral," depending on the specific emotions the model has been trained to recognize.

V. MODEL SELECTION AND TRAINING:

5.1 Objective: Train models capable of classifying emotions from audio input.

5.2 Techniques:

- **Convolutional Neural Networks (CNN):**
  - 2D CNN for Spectrograms: Treat the spectrograms as images, allowing CNNs to extract spatial features from the audio signal.
  - Use multiple convolutional layers, followed by pooling and fully connected layers to classify the emotion.
  - Popular libraries like TensorFlow or Keras can be used to implement CNN architectures.
- **Support Vector Machine (SVM):**
  - After extracting features (e.g., MFCCs or Spectrograms), use SVM for binary or multi-class classification.

- SVM is effective for smaller datasets where a deep learning model may not generalize well.
- **Random Forest:**
- Used for emotion classification based on extracted features.
- Random Forest works by constructing multiple decision trees and outputting the most common emotion prediction across trees.
- It provides high accuracy while being less computationally expensive compared to deep learning models.

VI. CONCLUSION:

The Speech Emotion Detection System Using Python is a new approach for recognizing human emotions from speech, which plays a central role in improving human-computer interaction. The use of machine learning and audio signal processing techniques enables this system to classify emotional states like happiness, sadness, anger, fear, surprise, and neutrality from raw speech input.Integration with advance feature extraction techniques like MFCC, pitch, energy, and spectral features gives it the critical patterns from which it is determined which one of the states has a particular emotion. Thus, the use of MLP also ensures accurate emotion classification using the extracted features, enabling reliable predictions to be built on the same extracted feature from the speech signal. This is also corroborated by the metrics obtained while evaluating accuracy, precision, recall, and F1-score, indicating how good the system performs in doing its task.The proposed system, with its flexibility and scalability, is aptly applied to various real-time scenarios such as customer service, mental health monitoring, and emotion-aware AI agents. Additionally, the ability of the system to process diverse speech data with different accents, noise levels, and languages makes it a robust solution for real-world applications.

Future improvements could be in the integration of more advanced deep learning techniques, such as CNNs or RNNs, to further enhance the system's ability to learn from raw audio data. Multimodal data, such as facial expressions and text, can also be incorporated to improve the accuracy and adaptability of the system in recognizing emotions. In conclusion, the speech emotion detection system presents massive promise in the future enhancement of human-computer interaction wherein machines are allowed to detect and respond to emotions related to humans and may well pave the way toward further empathetic applications by machine.

VII. ACKNOWLEDGMENT

REFERENCES

[1]Kumar, S., & Singh, A. (2022). Speech Emotion Recognition Using Deep Learning Models. *International Journal of Computer Applications*, 176(30), 1-5.
[2]Sharma, P., & Verma, R. (2021). Feature Extraction Techniques for Speech Emotion Recognition: A Review. *Journal of Artificial Intelligence Research*, 9(2), 45-56.
[3]Gupta, A., & Rani, P. (2023). Real-time Speech Emotion Detection System. *Proceedings of the Indian Conference on Speech and Signal Processing*, 112-117.
[4]Singh, R., & Kaur, M. (2022). An Overview of Speech Emotion Recognition in India. *International Journal of Machine Learning and Applications*, 8(4), 201-210.
[5]Tiwari, N., & Sharma, S. (2020). Advances in Deep Learning for Speech Emotion Recognition. *Journal of Data Science and Engineering*, 12(1), 34-42.
[6]Bhatia, S., & Gupta, K. (2021). Application of Machine Learning Algorithms in Emotion Recognition from Speech. *International Journal of Computational Intelligence Systems*, 14(3), 79-88.
[7]Patel, R., & Choudhary, N. (2022). Real-time Emotion Recognition through Audio Signals. *IEEE Conference on Emerging Technologies for Sustainable Development*, 145-150.
[8]Verma, P., & Yadav, A. (2020). A Comparative Study of Traditional and Deep Learning Approaches for Speech Emotion Detection. *International Journal of Artificial Intelligence and Soft Computing*, 6(2), 89-102.
[9]Bose, A., & Desai, V. (2023). Speech Emotion Recognition Using CNN and SVM: A Hybrid Model. *Proceedings of the Indian Society for Signal Processing and Applications*, 205-210.
[10]Sen, G., & Iyer, P. (2021). Feature-Based Approaches for Emotion Recognition from Speech. *International Journal of Speech and Signal Processing*, 15(1), 53-61.