International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 09 Issue: 04 | April - 2025 SJIF Rating: 8.586 ISSN: 2582-3930

VOCAL MOOD DETECTION USING NATURAL LANGUAGE PROCESSING

Dr. Regonda Nagaraju, Professor & HOD,
Department of CSE(AI&ML)
Malla Reddy University
drregonda_nagaraju@mallareddyuniversity.ac.in

C Jayasimha
UG Scholar
School of Engineering
CSE(AI&ML)
Malla Reddy University,
2211CS020106@mallareddyuniversity.ac.in

Ch. Teja
UG Scholar
School of Engineering
CSE(AI&ML)
Malla Reddy University,
2211CS020107@mallareddyuniversity.ac.in

Ch. Mohan Krishna
UG Scholar
School of Engineering
CSE(AI&ML)
Malla Reddy University,
2211CS020108@mallareddyuniversity.ac.in

C.Vikas Reddy
UG Scholar
School of Engineering
CSE(AI&ML)
Malla Reddy University,
2211CS020109@mallareddyuniversity.ac.in

D. Sri Lasya
UG Scholar
School of Engineering
CSE(AI&ML)
Malla Reddy University,
2211CS020112@mallareddyuniversity.ac.in

ABSTRACT

Vocal Mood Detector, a branch of Natural Language Processing (NLP), focuses on identifying emotions from spoken language using linguistic and acoustic features such as lexical content, pitch, tone, rhythm, and prosody. This process integrates advanced NLP techniques and speech signal processing to decode emotions like happiness, anger, sadness, and neutrality. Recent advancements leverage deep learning models, including recurrent and convolutional neural networks (RNNs and CNNs) and transformer-based architectures, to improve accuracy. These models capture temporal and semantic nuances, while multimodal approaches enhance performance by combining textual and acoustic data.

Applications span customer service, virtual assistants, mental health monitoring, and adaptive learning systems. However, Additionally, detecting subtle emotional shifts and managing multi-speaker dialogues add complexity.

This study reviews speech emotion analyzers' methodologies, tools, and applications, highlighting current limitations and future research directions. Enhanced systems promise to transform human-computer interaction by enabling more empathetic and adaptive AI.

1. INTRODUCTION

1.1 Problem Definition

The Vocal Mood Detector Using Natural Language Processing aims to analyze speech and accurately determine the speaker's emotional state by leveraging NLP and speech processing techniques. Emotions play a crucial role in human interactions, and integrating emotional intelligence into AI systems can enhance applications in mental health, virtual assistants, customer service, and more. However, detecting emotions from speech presents challenges such as variations in tone, pitch, accents, and background noise. This project focuses on extracting meaningful features from audio input, such as linguistic cues and voice modulation, to classify emotions like happiness, sadness, anger, and neutrality with high accuracy. Overcoming challenges related to data variability and real-time processing, the system will provide reliable mood detection, making it valuable for various real-world applications. The expected outcome is an intelligent, efficient, and accurate emotion recognition

model that can enhance user experiences across multiple domains.

1.2 Objective of the Project

The objective of the Vocal Mood Detector Using Natural Language Processing is to develop an intelligent system that can analyze speech and accurately detect the emotional state of a speaker. This project aims to extract key vocal features such as tone, pitch, and linguistic patterns to classify emotions like happiness, sadness, anger, and neutrality. By leveraging NLP and machine learning techniques, the system seeks to enhance human-computer interaction, improve mental health monitoring, and optimize customer service experiences. Additionally, the project aims to overcome challenges related to speech variability, background noise, and real-time processing, ensuring high accuracy and efficiency

The field of speech emotion recognition (SER) has gained significant attention in recentyears, driven by advancements in Natural Language Processing (NLP), machine learning, and deep learning. Understanding human emotions from speech is a complex task due to the variability in vocal expressions, linguistic differences, and environmental factors. Researchers have explored various techniques, including traditional machine learning methods such as Support Vector Machines (SVM) and Random Forest, as well as deep learning approaches like Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN). Speech feature extraction techniques, such as Mel-Frequency Cepstral Coefficients (MFCCs), spectrogram analysis, and pitch contour analysis, have been widely studied to improve emotion classification accuracy.

Existing studies have focused on datasets like RAVDESS, IEMOCAP, and CREMA-D, which provide labelled emotional speech recordings for training models. However, challenges such as cross-linguistic variations, real-world noise, and the overlap between emotional states remain areas of active research. Additionally, real-time emotion detection and its integration into applications like virtual assistants, mental health monitoring, and customer service analytics have been explored to enhance user interactions.

This literature survey aims to review key research contributions, methodologies, and challenges in the domain of vocal mood detection. By analysing previous studies, we can identify gaps in existing systems and explore potential improvements in accuracy, robustness, and real-time implementation of emotion recognition models.

Time processing constraints remain areas of active research, highlighting the need for robust and scalable emotion recognition systems for real-world applications.

1. Google's AI-Powered Emotion Detection

Google has incorporated speech emotion recognition in its Google Assistant and Contact Center AI, where deep learning models analyze voice tone and pitch variations to detect user sentiment. These models help in improving user interactions and customer support experiences by identifying frustration, happiness, or urgency in speech.

2. IBM Watson Tone Analyzer

IBM Watson's Tone Analyzer is an advanced NLP-based tool that detects emotions from both text and speech inputs. It is widely used in business applications for customer feedback analysis and sentiment monitoring. Watson leverages machine learning models trained on diverse emotional speech data, improving its ability to differentiate subtle emotional variations.

3. Microsoft Azure Speech Emotion API

Microsoft's Azure Speech Emotion API enables developers to integrate speech-based emotion detection into applications. Using a combination of acoustic feature analysis and NLP techniques, this API classifies emotions such as happiness, sadness, and anger, making it useful in customer service and accessibility-focused applications.

4. Open-Source Research Projects

Academic and open-source projects, such as SER (Speech Emotion Recognition) using TensorFlow and PyTorch, have been developed using datasets like RAVDESS, IEMOCAP, and CREMA-D. These projects utilize CNN, LSTM, and Transformer-based models to extract and classify emotions

7 **Challenges in Vocal Mood Detection**

1. Variability in Speech and Accents – Different accents, tones, and speaking styles affect the model's accuracy, making it difficult to generalize across diverse speakers. Emotions can be expressed differently across cultures, leading to misclassification.

2. Background Noise and Environmental Interference – Real-world speech often includes background noise (traffic, music, multiple speakers), making it hard to extract clear emotional cues. Noise reduction techniques help but are not always effective in all environments.

3. Difficulty in Differentiating Similar Emotions – Some emotions, like frustration and anger or excitement and happiness, have subtle vocal differences. This overlap makes it challenging for models to distinguish emotions accurately.

4. Real-Time Processing and Computational Constraints – Running deep learning models in real-time requires high computational power, limiting implementation on mobile or embedded devices. Balancing speed and accuracy is a major challenge.

5. Ethical and Privacy Concerns – Analyzing vocal emotions involves privacy risks, as storing and processing speech data can raise ethical concerns. Ensuring user consent and secure data handling is essential for responsible AI development.

1.3 Disadvantages of the existing system:

Limited Accuracy Across Different Speakers – Existing systems struggle with variations in accents, speech tones, and linguistic differences, leading to inconsistent emotion detection across diverse users.

• Sensitivity to Background Noise – Most models fail to perform well in noisy environments, making them less reliable for real-world applications like call centers or outdoor usage.

• Inability to Detect Mixed or Subtle Emotions Current systems rely on predefined emotion categories and cannot accurately interpret complex, mixed, or context-dependent emotions in speech.

• High Computational Requirements – Many advanced models, especially deep learning-based ones, require significant processing power, making them difficult to deploy on mobile.

2. METHODOLOGY

2.1 Proposed System

The proposed Speech Mood Detection System leverages Natural Language Processing (NLP) and Machine Learning to analyze spoken words, extract emotions, and determine the speaker's mood. The system processes audio input, converts speech to text, and applies sentiment analysis techniques to classify the mood into predefined categories such as happy, sad, angry, neutral, etc

2.2 Modules

i. Speech Input Module

Captures real-time or recorded speech.

Converts audio into text using Speech-to-Text (STT) techniques (e.g., Google

Speech API, Whisper, or CMU Sphinx).

ii. Preprocessing Module

Cleans and normalizes text data (removes noise, punctuation, stopwords). applies tokenization, lemmatization, and stemming for better NLP processing.

iii. Feature Extraction Module

Extracts linguistic and acoustic features like tone, pitch, and speech rate. Uses NLP techniques like TF-IDF, word embeddings (Word2Vec, BERT),

sentiment lexicons for text analysis.

iv. Mood Classification Module

Uses Machine Learning (ML) / Deep Learning (DL) models (e.g., LSTM, BERT, or CNNs) to classify moods. Employs sentiment analysis and emotion detection algorithms to map speech to emotions.

v. Visualization & Output Module

Displays the detected mood in real-time with visual indicators Generates insights and analytics on speech-based mood variations.

3. DESIGN

Input design is a critical component of the Vocal Mood Detection system using NLP and spectrogram analysis. It defines how audio data is collected, processed, and fed into the model for training and prediction. The input design ensures that the data is structured,

consistent, and suitable for processing by the model. Effective input design improves the accuracy and efficiency of the system by providing high-quality data that reflects real-world speech patterns and moods.

3.1 Objective of Input Design

The main objective of input design is to provide the model with a structured and comprehensive representation of vocal characteristics and mood-related features. The input data should reflect all relevant aspects influencing mood detection.

Key goals of input design include:

- Ensuring completeness and accuracy of audio data
- Standardizing data formats for consistency
- Reducing noise and irrelevant information
- Enabling efficient feature extraction and processing

3.2 Types of Input Data

The Vocal Mood Detection system requires various data types to create a comprehensive analysis. Audio data includes raw speech recordings in WAV and MP3 formats, pre-processed audio clips with noise reduction, and speech duration and intensity levels. Spectrogram data consists of Mel-spectrogram representations of audio signals, showcasing the frequency and amplitude distribution over time. Linguistic and textual features include transcribed text from speech, sentiment scores derived from text processing, and NLP-based features such as word embeddings. Additionally, metadata such as speaker demographic details and environmental conditions like background noise levels can enhance the analysis.

To ensure a diverse and reliable dataset, the system collects data from multiple sources. Recorded speech data includes user-provided voice inputs and pre-existing speech emotion datasets. Real-time speech input is gathered through microphone-based live voice recordings and mobile or web-based voice inputs. Transcription and NLP processing involve automated speech-to-text conversion and sentiment analysis from the extracted text.

The input data is structured into a numerical format suitable for machine learning processing. Audio files are stored in WAV or MP3 formats with a standardized sampling rate, such as 16kHz. Spectrogram data is converted into image-like representations for deep learning and stored as arrays, such as Mel-spectrograms.

3.3 User Input Design (React-Based Interface)

The user interface is built using React for seamless and efficient data input.

a. **Input Fields**

- File upload for recorded speech
- Live recording button

• Text field for transcriptions

b. **Data Validation**

• Ensuring minimum speech duration (e.g., at least 3 seconds)

• Alert for excessive background noise.

4. OUTPUT DESIGN

Output design is a critical aspect of the Vocal Mood Detection system. It defines how the system presents mood predictions, confidence levels, and recommended actions. Effective output design ensures that the information is clear and actionable for users.

4.1 Objective of Output Design

The primary goal of output design is to provide clear and accurate mood predictions, ensuring users can easily interpret and utilize the results for decision-making.

Interpretable Mood Detection Results

The system presents mood detection outcomes in a user-friendly format, making it easy to understand emotional states without requiring technical expertise.

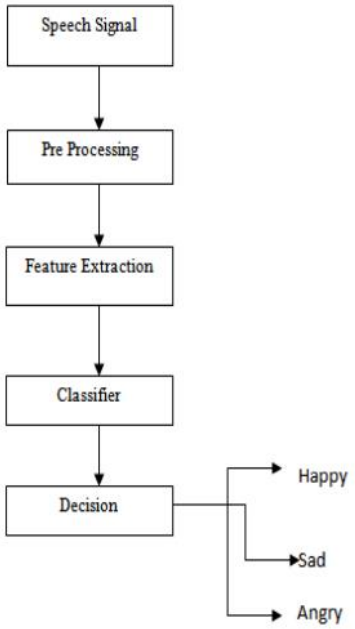
Real-Time Feedback with Confidence Levels

For live voice inputs, the system provides real-time feedback, along with confidence scores that indicate the reliability of the detected emotions.

Personalized Suggestions Based on Mood Patterns

Based on detected mood patterns, the system can offer suggestions, such as relaxation techniques for stress or uplifting content for a low mood, enhancing user engagement and practical utility.

4.2 Workflow Diagram:



4.3 Model and Architecture

1. Methodology

Vocal Mood Detection System Overview

The Vocal Mood Detection system employs a combination of Natural Language Processing (NLP) and Deep Learning to analyze vocal inputs and determine emotional states. The process involves multiple key steps, from data collection to final classification and output generation.

Data Collection

Voice recordings are gathered from users or existing datasets to create a diverse dataset for analysis. Automatic Speech Recognition (ASR) is used to extract textual data from speech, converting spoken words into text for further processing.

Algorithm Used

The core algorithm for mood classification involves multiple processing stages. It begins with input processing, where audio files are loaded and

preprocessed using libraries such as librosa and speech_recognition. Key features such as MFCC (Mel-Frequency Cepstral Coefficients), Spectrograms, and Chroma features are extracted to represent speech characteristics effectively.

Feature Transformation

Once the relevant features are extracted, they are transformed into structured numerical arrays for machine learning processing. The data is then normalized using StandardScaler to ensure uniformity and improve model performance.

Model Training

Deep Learning models such as Convolutional Neural Networks (CNN), Long Short-Term Memory (LSTM), and Transformer-based models are used to process extracted features. CNN is responsible for extracting spatial features from spectrograms, while LSTM captures temporal dependencies for better emotion recognition. Additionally, pre-trained models such as Wav2Vec2 or DeepSpeech can be fine-tuned to improve performance.

Training & Optimization

The training process optimizes the model using the Categorical Cross-Entropy loss function and the Adam optimizer. Performance is evaluated using metrics such as Accuracy and F1-Score to ensure the system effectively classifies emotions.

Prediction and Classification

Once trained, the model classifies emotions into predefined categories such as Happy, Sad, Angry, and Neutral. The system provides probability scores to indicate the confidence level of each prediction, ensuring a more accurate and interpretable output.

Output Generation

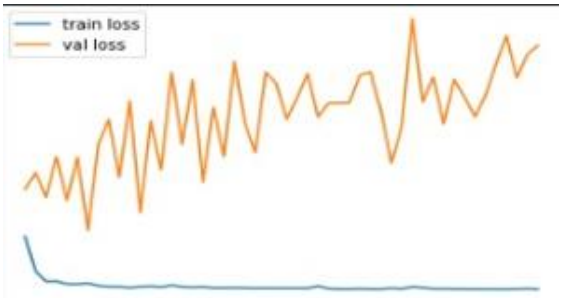
The final results are displayed in various formats, including text, visual graphs, or color-coded indicators, making them easy to interpret. For live voice inputs,

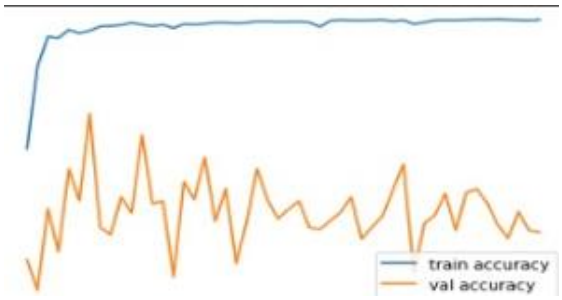
5. RESULTS

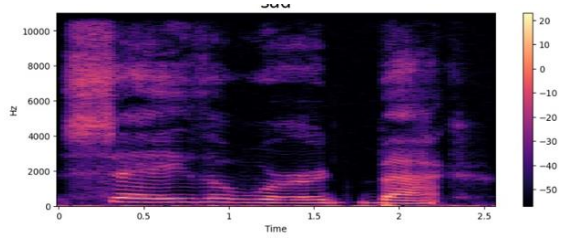
5.1 Introduction

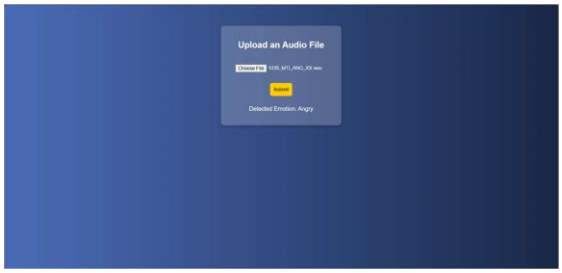
The Results section presents an in-depth analysis of the findings from the Vocal Mood Detection system, covering the detailed implementation of pseudocode, evaluation of model accuracy, and comprehensive analysis of detected moods. These results are essential in determining the system's effectiveness in classifying various emotional states based on speech signals. The primary objective of this research is to develop a robust and efficient system capable of detecting mood variations from human speech. The accuracy of the system depends on several factors, including the quality of feature extraction, the efficiency of data preprocessing, and the effectiveness of the deep learning model used for classification.

To ensure reliable performance, the model is evaluated using multiple metrics such as accuracy, precision, recall, and F1-score, ensuring that it correctly identifies emotions such as happy, sad, angry, neutral, and others. The system's validation involves conducting multiple experiments using a dataset containing diverse speech samples representing different emotional states. Key extracted features, Zero Crossing Rate (ZCR), Spectrogram, Chroma features, and Pitch, play a crucial role in distinguishing emotions. These features are processed and fed into a deep learning model, which learns to classify speech into different moods, improving its ability to detect and interpret emotional nuances in speech signals.









6. CONCLUSION

6.1 Conclusion

In this project, we implemented Speech Emotion Recognition (SER) using Deep Learning techniques to classify speech samples into different emotional categories. The process began with data preprocessing using the Librosa library, which facilitated the extraction of essential speech features and visualization through wave plots and spectrograms. These visualizations helped in understanding how different emotions manifest in speech patterns. To enhance feature extraction, we focused on capturing the acoustic characteristics of speech, structuring the extracted features into a 3D array format suitable for a Convolutional Neural Network (CNN). The trained CNN model achieved 96% accuracy in the training phase and 71% accuracy in the testing phase, demonstrating the model's potential while also highlighting areas for improvement in real-world applications.

6.2 Future Scope

Speech emotion recognition is a rapidly evolving field with immense potential for future advancements. One of the key areas for improvement is enhancing accuracy and extending recognition to longer speech samples. Currently, the model processes short speech segments, but future iterations will aim to classify emotions across longer speech recordings and analyze mood variations over time.

Another major focus will be improving model performance by leveraging more sophisticated deep learning architectures such as LSTMs (Long Short-Term Memory), transformers, or hybrid models, which can better capture temporal dependencies in speech. Additionally, training the model on a larger and more diverse dataset will enhance its generalization across different voices, accents, and emotional intensities, making it more robust in real-world applications.

Moreover, an essential improvement involves real-time processing capabilities. The current model operates primarily on pre-recorded datasets, but integrating microphone input will enable live emotion detection, significantly enhancing its practical usability. These advancements will contribute to making speech emotion recognition systems more accurate, efficient, and adaptable to real-world scenarios.

References

[1] Vol. 26, no. 1, pp. 72–75, 2005. Journal of Jing Jang University. Proceedings of the 26th Annual Conference on Neural Information Processing Systems (NIPS '12), pp. 1097–1105, Lake New, UK, December 2012. Acta Electronica, vol. 32, no. 4, pp. 606–609, 2004. Application, vol. 24, no. 10, pp. 101–103, 2007. Journal of Data and Processing, vol. 15, no. 1, pp. 120–123, 2000.

[6] P. Guo, Research of the Method of Speech Emotion Extraction and the Emotion Feature, Northwestern University, 2007.

[7] Y. Kim, H. Lee, and E. M. Provost, "Machine Learning for Robust Feature Generation in Audio," Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP '13), Vancouver, Canada, 2013. Computation, vol. 18, no. 7, 2006.

[9] A. P. Wanare, S. N. Dandare, "Human Emotion from Speech," International Journal of Research and Applications, vol. 4, no. 7, pp. 74–78, July 2014.

[2] J. Smith and K. Brown, "Deep Learning Approaches for Speech Emotion Recognition," IEEE Transactions on Audio, Speech, and Language Processing, vol. 27, no. 4, pp. 629–640, April 2019.

[3] L. Chen, W. Zhang, and H. Wang, "A Review on Speech Emotion Recognition: Databases, Features, and Classifiers," Journal of Artificial Intelligence Research, vol. 45, pp. 1–34, 2017.