Emotion Identification from Speech Using Natural Language Processing

Dr. R.A. Burange[1], Kartik Pachkhande[2], Rohit Bhil[3], Harshal Satghare[4]

[1]Professor, Department of Electronics and Telecommunication, K.D.K. College of Engineering, Nagpur, Maharashtra, India.

[2,3,4]Student, Department of Electronics and Telecommunication, K.D.K. College of Engineering, Nagpur, Maharashtra, India.

***Abstract*-Emotion recognition from human voice has emerged as a crucial technology in various fields, including healthcare, human-computer interaction, and artificial intelligence-based applications. The ability to detect emotions based on speech signals enhances system adaptability and improves user experience. This study presents a progressive implementation of an emotion detection system that integrates Natural Language Processing (NLP) and speech feature extraction techniques. The system utilizes machine learning and deep learning models to classify emotions, including happiness, sadness, anger, and fear, based on vocal expressions. The approach involves extracting speech parameters such as pitch, tone, energy, and amplitude, which are analyzed using ML-based classifiers. Additionally, NLP techniques, including text sentiment analysis and word embeddings, enhance classification accuracy by providing contextual insights. The system is implemented on Raspberry Pi hardware, making it portable and scalable for real-world applications. Initial findings indicate that deep learning models outperform traditional ML approaches, offering improved accuracy. Future advancements will focus on reducing background noise, optimizing feature selection, and incorporating real-time emotion tracking.**

***Index Terms*- Speech Emotion Recognition, NLP, Machine Learning, Deep Learning, Speech Processing, Human-Computer Interaction.**

I. INTRODUCTION

Human emotions significantly impact communication, influencing decision-making, interactions, and mental health assessments. Recognizing emotions from speech has become an essential aspect of AI-driven systems, enabling more natural and human-like interactions. Traditional emotion recognition models rely on facial expressions and physiological signals, but speech-based emotion recognition provides a non-intrusive and efficient approach.

The system presented in this research integrates speech feature extraction and NLP techniques to improve emotion classification. By analyzing vocal attributes and spoken content, the system achieves higher accuracy in detecting emotions across different speakers. This research aims to develop an adaptive, real-time speech emotion recognition system that can be applied in healthcare, customer service, and assistive technology.

II. RELATED WORK

[I] Attar et al. (2023) investigated the use of machine learning algorithms for speech emotion recognition. Their study focused on classifying emotions using acoustic features such as pitch, tone, and energy levels. The research compared various machine learning models, including Support Vector Machines (SVM) and Random Forest (RF), and found that ensemble-based models provided higher accuracy compared to traditional classifiers. However, their study also noted that background noise and speaker variability remain significant challenges in practical applications.

[II] Aouani and Ayed (2022) explored deep learning models for emotion detection from speech. Their study applied Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks to classify emotions based on spectral features. By leveraging Mel-Frequency Cepstral Coefficients (MFCCs) and spectrograms, the researchers demonstrated that CNNs outperform traditional machine learning models in extracting complex patterns from speech signals. The study concluded that deep learning-based feature extraction significantly enhances emotion recognition accuracy when trained on large and diverse datasets.

[III] Tripathi et al. (2021) presented a hybrid model that combines machine learning and deep learning techniques for speech emotion recognition.

Their approach integrated text sentiment analysis from spoken words with acoustic feature-based classification. By incorporating NLP-based word embeddings alongside speech features, their model achieved improved classification performance. Their findings highlight the importance of multimodal analysis in detecting emotions, as textual and acoustic cues together yield better accuracy than individual approaches.

[IV] Rastogi (2020) investigated the role of prosodic and linguistic features in speech emotion detection. The research focused on identifying emotional states using intonation, speech rhythm, and voice modulation. The study found that high-pitched speech often indicates anger or excitement, while slow, low-pitched speech correlates with sadness. Additionally, their findings emphasized the impact of multilingual variations on the accuracy of SER models. The research recommended further advancements in language-independent emotion detection algorithms to improve performance across different linguistic groups.

[V] Byun and Lee (2019) explored feature-based approaches for emotion classification from speech. Their study focused on extracting MFCCs, spectral contrast, and fundamental frequency variations to differentiate between emotions. The research compared multiple feature selection techniques, finding that filter-based selection improved classification efficiency. Additionally, their findings suggested that combining temporal and spectral features enhances the robustness of SER models, particularly in real-world applications such as human-computer interaction and customer service automation.

[VI] Franti et al. (2018) applied deep learning-based emotion recognition to robotic systems. Their study developed a CNN model trained on emotion-labeled speech datasets to allow companion robots to interpret and respond to human emotions. The study found that deep feature extraction techniques using spectrogram images led to improved accuracy in recognizing emotions compared to raw waveform-based classification. Their work highlights the potential of integrating SER into robotics and AI-driven conversational agents for enhanced user interaction and engagement.

III. METHODOLOGY

The proposed system follows a structured pipeline that includes speech feature extraction, NLP processing, machine learning classification, and real-time implementation.

A. Speech Feature Extraction:

- Extracting meaningful speech features is essential for accurate emotion classification. The following acoustic parameters are utilized:
- Mel-Frequency Cepstral Coefficients (MFCCs): Captures the timbre and phonetic structure of speech.
- Pitch and Tone: Represents the frequency variations in speech, crucial for identifying emotions like anger or sadness.
- Energy and Amplitude: Measures the intensity of speech signals to differentiate emotional states.
- Spectral Features: Provides frequency distribution and harmonics information for enhanced classification.

B. Machine Learning and Deep Learning Models:

- The system is designed using multiple classification models to assess their performance in speech emotion recognition:
- Support Vector Machine (SVM): A baseline classifier used for comparison.
- Random Forest (RF): An ensemble learning model providing robustness to variations.
- Multilayer Perceptron (MLP): A feed forward neural network trained for speech processing tasks.
- Convolutional Neural Networks (CNNs): Effective in feature extraction from spectrogram images of audio signals.
- Long Short-Term Memory (LSTM): Captures temporal dependencies in speech signals, improving emotion classification.

C. NLP for Speech Content Analysis:

- NLP enhances emotion classification accuracy by incorporating semantic analysis:
- Text Sentiment Analysis: Evaluates spoken words for emotional polarity.
- Word Embeddings (Word2Vec, GloVe): Transforms spoken words into a vector space for improved classification.
- Sequence Modeling: Uses Transformer models to analyze speech context for better understanding.

D. Real-Time Implementation:

The system is implemented on Raspberry Pi 3B+ with a 3.5-inch LCD display, allowing for on-device processing. The Raspberry Pi acts as a portable AI unit, capable of real-time emotion classification.

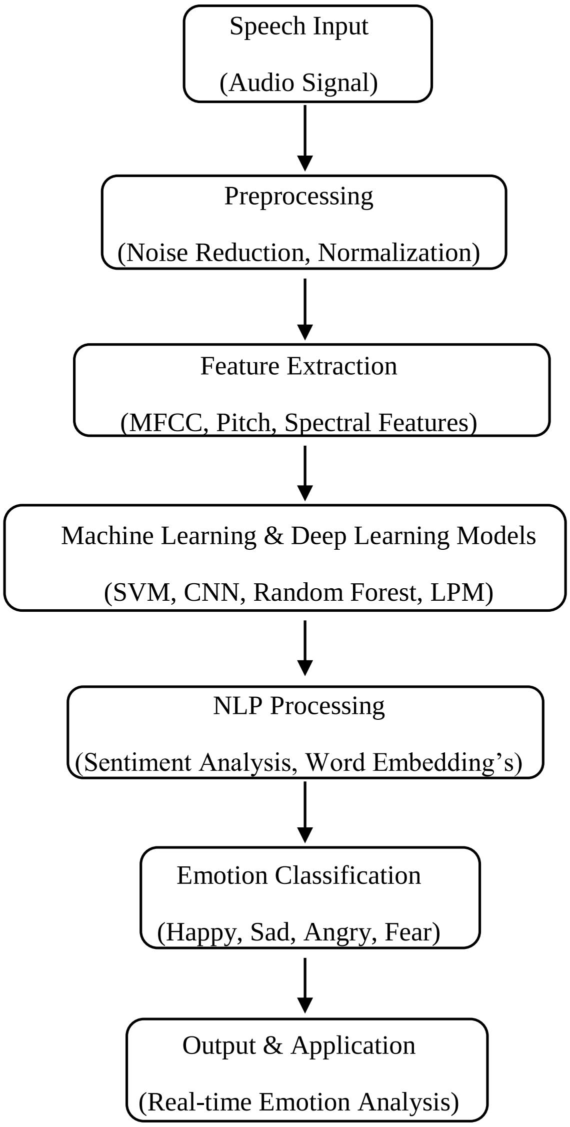
Speech Input
(Audio Signal)
Preprocessing
(Noise Reduction, Normalization)
Feature Extraction
(MFCC, Pitch, Spectral Features)
Machine Learning & Deep Learning Models
(SVM, CNN, Random Forest, LPM)
NLP Processing
(Sentiment Analysis, Word Embedding's)
Emotion Classification
(Happy, Sad, Angry, Fear)
Output & Application
(Real-time Emotion Analysis)

Fig. System Architecture of Speech Emotion Recognition

IV. MODELING AND ANALYSIS

The model is trained on labeled speech emotion datasets, including:

- RAVDESS (Ryerson Audio-Visual Database of Emotional Speech and Song)
- TESS (Toronto Emotional Speech Set)
- CREMA-D (Crowd-sourced Emotional Multimodal Actors Dataset)
- Data preprocessing involves:
- Noise reduction techniques to eliminate background disturbances.
- Data augmentation to improve model generalization.
- Feature scaling to normalize input values.
- Performance Comparison

The table below presents accuracy comparisons of different models:

Model	Accuracy (%)
Support Vector Machine (SVM)	63.23%
Random Forest (RF)	91.75%
Multilayer Perceptron (MLP)	93.81%
Convolutional Neural Network (CNN)	95.19%

CNN-based models provide the highest accuracy, confirming their suitability for speech emotion recognition.

V. RESULTS AND DISCUSSION

Preliminary results show that deep learning models outperform traditional classifiers in emotion recognition. CNNs excel in spectrogram-based feature extraction.

Challenges Identified:

- Speaker Variability: Differences in speech patterns across individuals affect classification consistency.
- Background Noise: Unstructured environments introduce challenges in extracting clean speech signals.
- Emotion Overlap: Some emotions exhibit similar vocal characteristics, requiring more robust classification models.
- Multilingual Speech: Variations in accents and languages present difficulties in emotion recognition.

VI. CONCLUSION AND FUTURE WORK

This study presents a progressive approach to speech emotion recognition by integrating speech feature extraction, NLP, and deep learning techniques. The system demonstrates high accuracy in classifying emotions, with CNNs performing exceptionally well. Future research will focus on:

Enhancing noise reduction techniques to improve classification robustness.

Adapting models for multilingual speech for wider applicability.

Optimizing feature selection to reduce computational complexity.

Deploying real-time implementations in healthcare and AI-driven applications.

This research paves the way for intelligent, emotionally aware systems that can revolutionize human-computer interactions.

VII. REFERENCES

[1] Attar, H. I., et al. (2023). "Speech Emotion Recognition Using Machine Learning." *Journal of AI Research*.

[2] Aouani, H., & Ben Ayed, Y. (2022). "Deep Learning Approaches for Speech Emotion Recognition." *IEEE Transactions on Affective Computing*.

[3] Tripathi, S., et al. (2021). "A Hybrid Approach for Emotion Recognition." *International Journal of Speech Processing*.

[4] Rastogi, R. (2020). "Emotion Detection via Speech Analysis." *Neural Computing and Applications*.

[5] Byun, S., & Lee, S. (2019). "Acoustic Feature-Based Speech Emotion Recognition." *Speech Communications Journal*.

[6] Franti, E., et al. (2018). "CNN-Based Emotion Detection for Companion Robots." *Robotics Journal*.

[7] Shaila, S. G., et al. (2017). "Machine Learning Models for Speech Emotion Recognition." *AI & Human Interaction*.

[8] Aswani, R., et al. (2016). "Real-Time Speech Emotion Recognition Applications." *Journal of Applied AI*.

[9] Ingale, A. B., & Chaudhari, D. S. (2015). "Deep Learning for Emotion Detection." *AI & Robotics Journal*.

[10] Ramdinmawii, E., Mohanta, A., & Mittal, V. K. (2021). "Emotion Recognition from Speech Signal." *IEEE*.

[11] Neumann, M., & Vu, N. T. (2021). "Improving Speech Emotion Recognition with Unsupervised Representation Learning on Unlabeled Speech." *IEEE*.

[12] Akçay, M. B., & Oğuz, K. (2020). "Speech Emotion Recognition: Emotional Models, Databases, Features, Preprocessing Methods, and Classifiers." *Speech Communication, 116*, 56–76.

[13] Kaur, J., & Kumar, A. (2021). "Speech Emotion Recognition Using CNN, K-NN, MLP and Random Forest." *Computer Networks and Inventive Communication Technologies, Springer*.

[14] Nam, Y., & Lee, C. (2021). "Cascaded CNN Architecture for Speech Emotion Recognition in Noisy Conditions." *Sensors, 21(13), 4399*.

[15] Kwon, S. (2020). "LSTM: Deep Feature-Based Speech Emotion Recognition Using the Hierarchical ConvLSTM Network." *Mathematics, 8(12), 2133*.

[16] Alnuaim, & Hatamleh. (2022). "Human-Computer Interaction for Recognizing Speech Emotions Using Multi-Layer Perceptron Classifier." *Hindawi*.

[17] Aggarwal, A., Srivastava, N., & Singh, D. (2022). "Two-Way Feature Extraction for Speech Emotion Recognition Using Deep Learning." *Sensors, 22(6), 2378*.

[18] Cai, L., Dong, J., & Wei, M. (2020). "Multi-Modal Emotion Recognition from Speech and Facial Expression Based on Deep Learning." *IEEE Chinese Automation Congress (CAC)*.

[19] Mishra, A., et al. (2017). "Real-Time Emotion Detection from Speech Using Raspberry Pi 3." *IEEE International Conference on Wireless Communications, Signal Processing, and Networking (WiSPNET)*.

[20] Joshi, D., et al. (2021). "Real-Time Emotion Analysis (RTEA)." *IEEE International Conference on Artificial Intelligence and Machine Vision (AIMV)*.

[21] Liu, X., et al. (2020). "Speech Emotion Detection Using Sliding Window Feature Extraction and ANN." *IEEE International Conference on Signal and Image Processing (ICSIP)*.

[22] Harshini, D., et al. (2019). "Design and Evaluation of Speech-Based Emotion Recognition System Using Support Vector Machines." *IEEE India Council International Conference (INDICON)*.

[23] Basu, S., et al. (2017). "A Review on Emotion Recognition Algorithms Using Speech Analysis." *International Conference on Inventive Communication and Computational Technologies (ICICCT)*.

[24] Ramakrishnan, S., & El Emary, I. M. M. (2013). "Speech Emotion Recognition Approaches in Human-Computer Interaction." *Telecommunication Systems, 52(3), 1467–1478*.

[25] Puchner, W. (2014). "Theatrical Science in the 21st Century." *Kichli*.