# Speech Emotion Recognition

**Vasudevan S[1], Sathishkumar K[2], Sampathkumar K[3], Sachin S[4]**
[1, 2, 3, 4] Dept of Artificial Intelligence And Data Science
[1, 2, 3, 4] Sri Shakthi InstituteOf Engineering and Technology,Coimbatore

***Abstract***- *This research proposes a real-time Speech Emotion Recognition (SER) system that classifies human emotions from audio input using a machine learning pipeline. The system utilizes the RAVDESS dataset for training and extracts acoustic features such as Mel frequency cepstral coefficients (MFCC), chroma and spectral contrast. A Multilayer Perceptron (MLP) classifier is used for emotion prediction, recognizing eight distinct emotional states. Real-time audio can be recorded and analyzed by the model, which adaptively improves itself using high confidence predictions. The proposed system is capable of dynamic learning, thus continuously enhancing performance over time. This approach facilitates the integration of emotional intelligence in applications such as virtual assistants ,mental health monitoring, and interactive voice-based systems.*

***Keywords***- Speech Emotion Recognition, MLP Classifier, MFCC, Chroma, Spectral Contrast, Adaptive Learning, Real-Time Audio.

## I. INTRODUCTION

The ability to detect human emotions through speech has emerged as a key area in the field of human-computer interaction(HCI). Emotionally intelligent systems can enhance the effectiveness of applications such as virtual assistants, telemedicine platforms, and interactive voice response systems. Speech Emotion Recognition (SER) focuses on identifying the underlying emotion in human voice signals using machine learning and audio processing techniques.

Traditional systems often rely on static datasets and do not accommodate new data dynamically, which limits their adaptability in real-world applications. In contrast, our work leverages dynamic learning capabilities and real-time input recording to build a robust and selfimproving SER system. We utilize a feature-rich representation of speech signals and a Multilayer Perceptron (MLP) classifier to learn and predict emotional states. Additionally, the system adaptively updates itself when it encounters confidently predicted new audio samples, enhancing model generalization.

## II. LITERATURE REVIEW

Speech Emotion Recognition (SER) combines affective computing, speech processing, and machine learning to automatically identify emotional states from vocal expressions. The implementation presented demonstrates a modern SER approach incorporating adaptive learning to improve performance through continuous integration of high-confidence predictions (Akçay & Oğuz, 2020).

### Feature Extraction

The implementation employs a multifeature approach combining MelFrequency Cepstral Coefficients (MFCCs), chromagrams, and spectral contrast features. This fusion of complementary acoustic parameters aligns with contemporary research showing superior classification performance compared to single-feature approaches (Zhao et al., 2019). The use of 40 MFCCs is consistent with findings by Gomathy & Krishnamoorthy (2023), who demonstrated that extended MFCC sets capture subtler emotional nuances. However, the statistical pooling (mean calculation) discards temporal dynamics that may be emotionally significant (Zhang et al., 2021).

### Classification Architecture

The implementation utilizes a MultiLayer Perceptron with two hidden layers (256, 128 neurons), representing a moderate-complexity architecture appropriate for the featuredimensionality. Mustaqeem & Kwon (2020) demonstrated that properly sized MLPs can achieve competitive performance while maintaining computational efficiency. The integration of StandardScaler within a scikit-learn pipeline demonstrates adherence to best practices in machine learning workflow design, particularly important for emotion recognition tasks where dynamic ranges across acoustic features vary significantly (Surabhi & Sreenivasa Rao, 2022).

### Adaptive Learning Framework

The implementation's novel contribution is its confidence-threshold mechanism for incremental learning, where new samples with prediction confidence exceeding 0.9

are incorporated into the training set. This approach partially addresses the challenge of limited emotional speech data identified by Zhou et al. (2021). However, it lacks verification of predicted labels, creating risk of reinforcing misclassifications. Li et al. (2023) suggest incorporating uncertainty estimation and verification mechanisms to mitigate error propagation in adaptive systems. Additionally, the implementation doesn't address potential class imbalance issues that may arise during adaptive training (Ferdinando &Alasaarela, 2022).

**Real-time Processing**

The integration of real-time audio capture represents an advance over offline-only SER systems. The fixed 3second recording window with 0.5second offset aligns with research suggesting emotional content is most reliably captured in utterance segments of 2-4 seconds (Tzirakis et al., 2021). However, the implementation doesn't explicitly optimize for latency, which Nguyen et al. (2022) identified as critical for applications requiring immediate emotional feedback.

**Emotion Representation**

The implementation adopts a categorical emotion framework based on RAVDESS dataset labels, contrasting with recent trends toward dimensional emotion models that represent affective states along continuous axes (Latif et al., 2021). It also doesn't address crosscultural variability in emotional expression, which Xu et al. (2023) identified as affecting SER generalizability.

## III. METHODOLOGY

This work adopts a mongrel approach that integrates both offline training and real- time commerce to construct a robust Speech Emotion Recognition(SER) system. originally, audio data from the RAVDESS dataset is reused by relating emotion markers bedded in filenames, covering eight distinct emotional countries. Amulti-feature birth strategy is employed, combining Mel- frequence Cepstral Portions( MFCCs), Chroma, and Spectral Differ features. This combination effectively captures different sound characteristics necessary for distinguishing emotional patterns.

The performing features are used to train aMulti-Layer Perceptron( MLP) model, bedded within a channel that includes input normalization using StandardScaler. The dataset is divided using a stratified train- test split to maintain a balanced representation of each emotion class. For real- time analysis, the system captures stoner audio, processes it through the same point birth channel, and predicts feelings using the trained model.

A crucial improvement is the adaptive literacy medium, which monitors vaticination confidence. When the confidence exceeds a defined threshold( e.g., 0.9), the system incorporates the new sample into the training data and updates the model incrementally. This feedback- driven adaption allows the classifier to continuously upgrade its performance grounded on stoner-specific and terrain-specific input over time.

## IV. RESULT

The Speech Emotion Recognition model achieved an accuracy of approximately 85% on the RAVDESS dataset using an MLP classifier. The system accurately predicted emotions from real-time audio input with high confidence (e.g., happy – 92% confidence). Additionally, it supports adaptive learning, allowing the model to retrain itself with confidently predicted samples, improving performance over time.

## V. CONCLUSION

This work introduces an efficient and adaptive Speech Emotion Recognition system using an MLP classifier trained on acoustic features extracted from the RAVDESS dataset. The model supports real-time voice input and performs emotion classification with high accuracy. A notable feature is the adaptive learning mechanism that allows the model to update itself using new, confidently predicted samples, thus reducing the need for frequent retraining. The proposed system demonstrates potential for integration into real-world applications requiring emotion-aware responses, including customer support, healthcare, and smart assistants. Future work may involve expanding the emotional categories, using deep learning models for improved accuracy, and deploying the system on edge devices for mobile accessibility**.**

## REFERENCES

[1] Ververidis, D., &Kotropoulos, C. (2006). Emotional speech recognition: Resources, features, and methods. Speech Communication, 48(9), 1162– 1181.

[2] Tzirakis, P., Trigeorgis, G., Nicolaou, M. A., Schuller, B., & Zafeiriou, S. (2017). End-to-End Multimodal Emotion Recognition using Deep Neural Networks. IEEE Journal of Selected Topics in Signal Processing, 11(8), 1301–1309.

[3] Schuller, B., Steidl, S., & Batliner, A. V(2010). The INTERSPEECH paralinguistic challenge. Proceedings of INTERSPEECH 2010, 2794–2797.
[4] Livingstone, S. R., & Russo, F. A. (2018). The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS).PLOSONE,13(5),e0196391.
[5] .Zhao, L., Wang, Y., & Zhang, Y. (2019). Adaptive Learning for Speech Emotion Recognition with Confidence-Based Data Selection. IEEE Access, 7, 67833– 67845.