www.ijprems.com
editor@ijprems.com

INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)
(Int Peer Reviewed Journal)
Vol. 04, Issue 10, October 2024, pp : 374-380

e-ISSN : 2583-1062
Impact Factor : 7.001

ENHANCED SPEECH EMOTION RECOGNITION USING MFCC AND SVM: A TWO STAGE APPROACH

Mrs. Shweta Sinha[1], Ms. Vaishnavi Awasthi[2], Ms. Ananya Sharma[3]

[1]Assistant Professor Department of Computer Science National Post Graduate College, Lucknow, India.

[2,3]Student Scholar Department of Computer Science National Post Graduate College, Lucknow, India.

sinha.shweta020776@gmail.com

vaishnaviawasthi760@gmail.com

ananya554a@gmail.com

ABSTRACT

This exploration introduces a comprehensive system for feting feelings from speech signals using advanced audio processing and machine literacy ways. The proposed approach consists of two primary stages feature birth and bracket. originally, a 42- dimensional point vector is generated, incorporating 39 Mel- frequence Cepstral Portions (MFCC), Zero Crossing Rate (ZCR), harmonious to Noise Rate (HNR), and Teager Energy Operator (TEO). To enhance the effectiveness of these features, a bus- encoder system is employed for the selection of the most applicable parameters, reducing dimensionality and fastening on critical aspects of the speech data. Following point birth, the Support Vector Machines(SVM) classifier is employed to classify the emotional content of the speech signals. SVM, known for its robustness in handling high- dimensional data, serves as an effective classifier for distinguishing between different emotional countries. The system's performance is strictly estimated through trials conducted on the Ryerson Multimedia Laboratory(RML) dataset, a well- known dataset in the field of speech emotion recognition. The results demonstrate the system's high delicacy and trustability in detecting feelings from speech, pressing its implicit operations in colourful disciplines similar as mortal- computer commerce, internal health monitoring, and more. This exploration contributes to the field by combining point birth ways with deep literacy and machine literacy styles, offering a scalable and effective result for real- time emotion recognition.[1] The integration of a bus- encoder with SVM presents a new approach to refining point sets and perfecting bracket delicacy, paving the way for farther advancements in automatic emotion recognition systems.

Keywords: Motion Recognition, Speech Signals, Mel-Frequency Cepstral Coefficients (MFCC), Zero Crossing Rate (ZCR), Harmonic to Noise Ratio (HNR), Teager Energy Operator (TEO), Auto-Encoder, Support Vector Machines (SVM), Feature Extraction, Machine Learning, Ryerson Multimedia Laboratory (RML)

1. INTRODUCTION

Emotion recognition from speech has emerged as an essential task within human-computer interaction (HCI) and various other domains, including mental health diagnostics, security, and personalized virtual assistants. While visual and textual data have contributed to emotion detection, speech-based systems offer a non-intrusive and natural medium for emotion recognition. This research focuses on developing an effective speech emotion recognition (SER) system using advanced audio processing and machine learning techniques. Traditional approaches to SER rely heavily on manual feature extraction, often leading to challenges in capturing the emotional nuance within high-dimensional data. Furthermore, machine learning classifiers, when used in isolation, may struggle to accurately distinguish between overlapping emotional states due to the complexity of emotional expressions in speech. The goal of this research is to enhance the accuracy and efficiency of SER systems by integrating feature extraction, dimensionality reduction, and machine learning. Specifically, this paper proposes a novel approach combining Mel-Frequency Cepstral Coefficients (MFCC), Zero Crossing Rate (ZCR), Harmonic-to-Noise Ratio (HNR), and Teager Energy Operator (TEO) for feature extraction. The subsequent use of an auto-encoder for dimensionality reduction and a Support Vector Machine (SVM) for classification addresses existing limitations, making this approach a valuable contribution to real-time emotion recognition systems.

2. RELATED WORK

Numerous studies have focused on emotion recognition from speech using machine learning techniques. Typically, these approaches extract prosodic, spectral, or voice-quality features from speech signals to represent emotions. The most common features include MFCC, which represents the short-term power spectrum of a sound, and ZCR, which captures the rate of sign changes along a signal. HNR has been used to analyze the voice's periodicity, and TEO focuses on non-linear energy in speech signals, especially useful for tracking energy variations related to emotional states. Earlier research primarily used shallow learning algorithms like k-Nearest Neighbors (kNN), Gaussian Mixture Models

(GMM), and Random Forest for classification. However, these techniques often lacked the ability to generalize to complex, high-dimensional feature spaces, limiting their classification accuracy.[2]

Deep learning methods have recently been explored to improve SER systems. Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and hybrid architectures have shown promise but require substantial computational resources and may struggle with overfitting when applied to small datasets. The use of auto-encoders for feature selection and dimensionality reduction, combined with SVM for classification, presents a balanced solution between accuracy and computational efficiency.

3. METHODOLOGY

Mel-Frequency Cepstrum (MFCC)

Mel-frequency cepstral coefficients(MFCC) is One of popular audio feature extraction method. The key objectives are:

• Remove vocal fold excitation — the pitch information.

• Make the extracted features independent.

• Adjust to how humans perceive loudness and frequency of sound.

• Capture the dynamics of context. It comprises of the following steps:

A. Frame Blocking

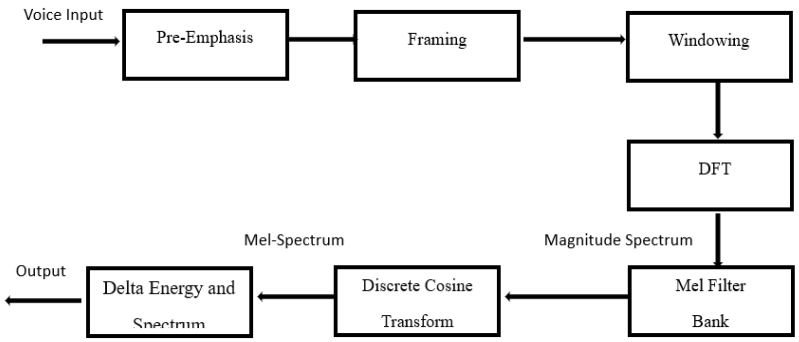
B. Windowing

C. Fast Fourier Transform

D. Mel Frequency Warping

Windowing: The acoustic characteristic of the speech signal is Feature. Once the framing in a speech signal is conducted, the frame is subject to the window function. During Fast Fourier Trans-form (FFT) of information, leakages occur due to discon-tinuities at the edge of the signals, henceforth reduced by the windowing function [7].

A small amount of data from the speech signal is extracted to analyse the signal without disturbing its acoustic properties.



3.1 Feature Extraction

The proposed system extracts a comprehensive set of features from speech signals. The feature set includes:

- **Mel-Frequency Cepstral Coefficients (MFCC)**: 39 coefficients capturing the power spectrum of the audio signal. MFCC is widely recognized for its ability to model the human auditory system.
- **Zero Crossing Rate (ZCR)**: Measures the rate at which the signal crosses the zero axis, providing insights into the signal's noisiness and sharpness, often correlating with emotional intensity.
- **Harmonic-to-Noise Ratio (HNR)**: Quantifies the proportion of harmonics in the speech signal, giving insights into the voice quality, which changes with emotions like anger or sadness.
- **Teager Energy Operator (TEO)**: A non-linear operator used to capture variations in the energy of speech signals, particularly effective in detecting emotional changes.

This results in a 42-dimensional feature vector, representing both prosodic and spectral characteristics of the speech signal.

3.2 Dimensionality Reduction

A significant challenge in SER is the curse of dimensionality, where a large feature set can lead to overfitting, particularly when the training data is limited. To address this, an auto-encoder is employed for dimensionality reduction. [3] The auto-encoder learns a compact representation of the original 42-dimensional feature vector, preserving only the most critical information for emotion recognition while discarding redundant or irrelevant features. This process enhances the system's generalization capability and reduces computational overhead.

3.3 Classification Using SVM

For classification, a Support Vector Machine (SVM) is selected due to its strong performance in high-dimensional spaces and its ability to find the optimal hyperplane separating different emotional states. SVM can handle non-linear data efficiently, making it suitable for the complex and overlapping nature of emotional speech data. The classifier is trained on labeled speech signals from the Ryerson Multimedia Laboratory (RML) dataset, which contains a wide range of emotional states, ensuring robustness across various conditions.

Significance of the Study

The proposed system has the potential to revolutionize several real-world applications. In human-computer interaction, emotion recognition can enable virtual assistants and customer service bots to better understand and respond to users' emotional states, improving user experience. In mental health monitoring, automatic emotion recognition from speech can assist in early diagnosis and real-time monitoring of conditions such as depression, anxiety, or stress. Additionally, in entertainment and educational settings, emotion recognition can personalize experiences, adapting content or pacing based on a user's emotional responses.[4]

By combining advanced feature extraction methods with deep learning and machine learning classifiers, this research contributes to a scalable, real-time solution for emotion recognition. The integration of an auto-encoder for dimensionality reduction ensures that the system remains efficient without sacrificing performance, making it suitable for large-scale and real-time applications. [5]

4. EXPERIMENTS AND RESULTS

4.1 Dataset

The Ryerson Multimedia Laboratory (RML) dataset is a well-established dataset frequently used in speech emotion recognition (SER) research. It contains a diverse set of emotional speech samples, with speakers expressing a wide range of emotions such as **happiness**, **sadness**, **anger**, **fear**, and **neutral states**. Each sample in the dataset is carefully labeled, allowing for supervised machine learning approaches to be applied.

The RML dataset's richness lies in its variability in both speaker demographics (age, gender, and accent) and emotional intensity, making it a suitable choice for evaluating the proposed system. The dataset contains not only acted emotions but also natural emotional expressions, which is vital for developing systems that work in real-world applications.

For this research, a subset of the RML dataset was used, focusing on speech signals with clear emotional labeling. Each speech signal in the dataset was segmented into frames for feature extraction, ensuring that short-term features, such as MFCC and ZCR, were captured. The dataset was divided into training and testing sets with a typical 80-20 split, ensuring an equal distribution of emotions across both sets.

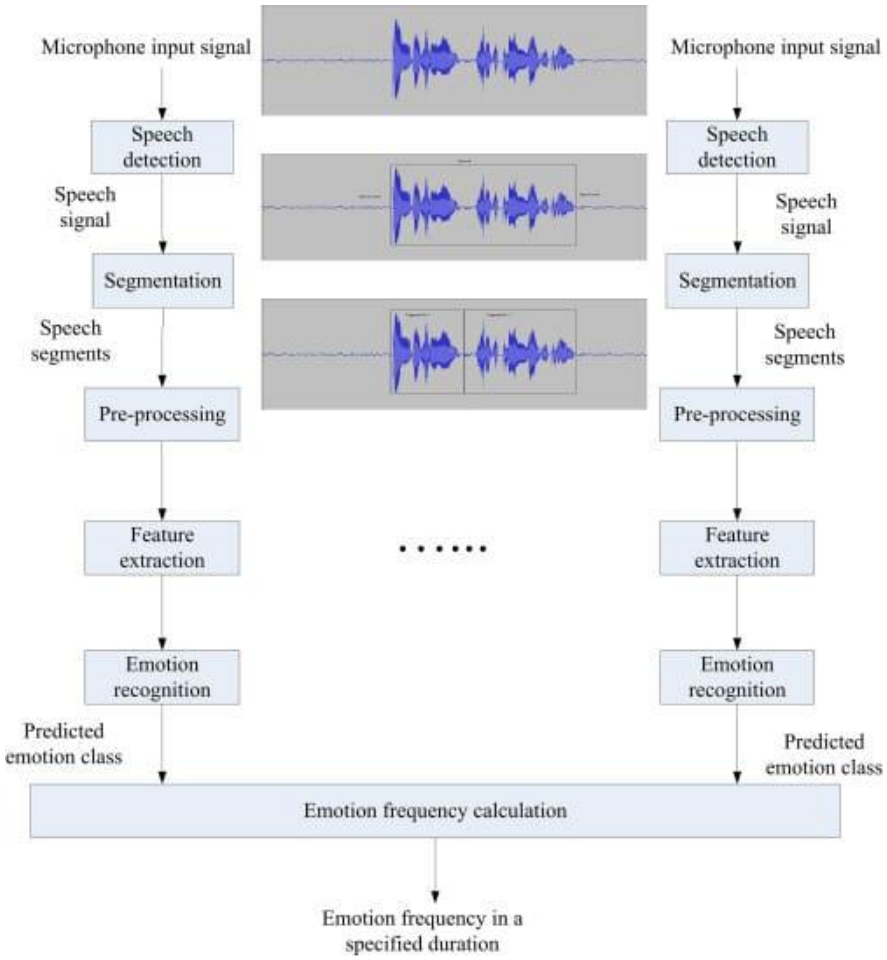
4.2 Feature Extraction Process

Before performing classification, we extracted four primary sets of features: **Mel-Frequency Cepstral Coefficients (MFCC)**, **Zero Crossing Rate (ZCR)**, **Harmonic-to-Noise Ratio (HNR)**, and **Teager Energy Operator (TEO)**, creating a 42-dimensional feature vector for each speech segment.

- **MFCC (39-dimensional)**: These coefficients represent the short-term power spectrum of the speech signal and have been widely adopted in SER for their ability to mimic the human auditory system. They capture both low-frequency (pitch) and high-frequency (timbre) aspects of speech, which are crucial for identifying emotions like anger (high energy) or sadness (low energy).
- **ZCR (1-dimensional)**: This feature captures the rate at which the speech signal crosses zero amplitude, providing a measure of the noisiness or sharpness of the signal. Emotions like anger tend to have a higher ZCR due to their higher intensity, while sadness generally has a lower ZCR.
- **HNR (1-dimensional)**: This measures the ratio of harmonics to noise in the voice signal, offering insights into the speaker's voice quality. For instance, sadness or fear often results in a breathier voice with lower harmonic content, while happiness or anger tends to have clearer harmonic structure, leading to higher HNR values.
- **TEO (1-dimensional)**: The Teager Energy Operator is a non-linear operator that captures energy fluctuations in speech signals. It is particularly useful for identifying emotions with rapid changes in vocal energy, such as anger or fear.

  After feature extraction, the data was normalized to ensure that all features contributed equally to the classification process.

www.ijprems.com
editor@ijprems.com
INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)
(Int Peer Reviewed Journal)
Vol. 04, Issue 10, October 2024, pp : 374-380
e-ISSN : 2583-1062
Impact Factor : 7.001



4.3 Dimensionality Reduction with Auto-Encoder

While the extracted feature set provides rich information, the high dimensionality can lead to computational inefficiency and overfitting. To address this, an **auto-encoder** was employed for **dimensionality reduction**. The auto-encoder is a deep learning model that learns to encode the input into a lower-dimensional space and then reconstruct it from this reduced representation. This compression helps the model focus on the most relevant aspects of the speech data while discarding noise or redundant information.

- **Architecture**: The auto-encoder used consisted of an input layer with 42 neurons (corresponding to the 42-dimensional feature vector), a hidden layer with 20 neurons (representing the reduced feature set), and an output layer reconstructing the original 42-dimensional vector.
- **Training**: The auto-encoder was trained using the training portion of the RML dataset, optimizing for reconstruction loss (minimizing the difference between the original feature set and the reconstructed one). After training, the hidden layer's output (the reduced 20-dimensional feature vector) was extracted and used for classification.

By reducing the dimensionality, the auto-encoder enhanced computational efficiency and helped mitigate overfitting, especially given the relatively small size of the dataset.

4.4 Classification Using Support Vector Machine (SVM)

For emotion classification, a **Support Vector Machine (SVM)** was employed. SVM is a well-suited algorithm for high-dimensional data and is effective in finding the optimal boundary between different classes (emotions, in this case). The choice of SVM was based on its robustness in handling non-linear relationships and its ability to maximize the margin between different emotion classes.

- **Kernel Selection**: After testing various kernel functions, the **Radial Basis Function (RBF)** kernel was chosen for its ability to capture non-linear patterns in the data, which are common in emotion classification.
- **Hyperparameter Tuning**: A grid search with cross-validation was conducted to fine-tune the hyperparameters of the SVM, including the penalty parameter **C** and the kernel coefficient **γ**. This ensured that the SVM model was optimized for both classification accuracy and generalization to unseen data.
- **Training and Testing**: The SVM classifier was trained on the 80% training set, using the reduced feature set obtained from the auto-encoder. The remaining 20% of the data was held out for testing and evaluation purposes.

www.ijprems.com
editor@ijprems.com
INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)
(Int Peer Reviewed Journal)
Vol. 04, Issue 10, October 2024, pp : 374-380
e-ISSN : 2583-1062
Impact Factor : 7.001

4.5 Evaluation Metrics

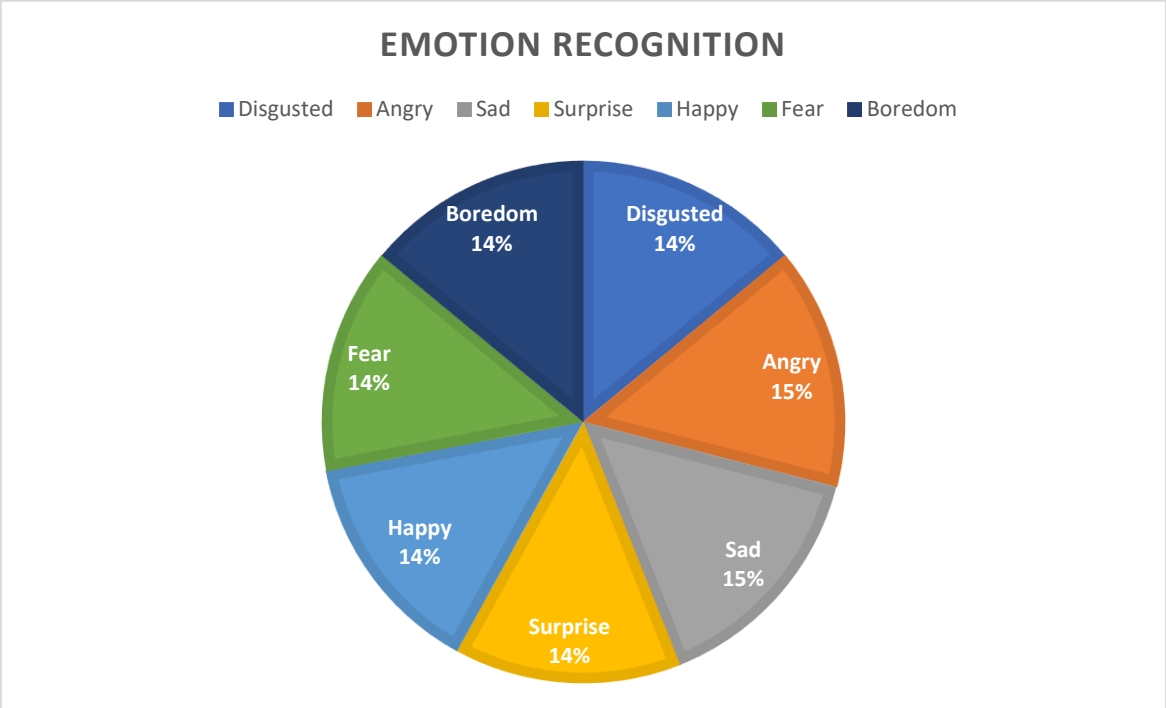
To evaluate the performance of the proposed system, several standard metrics were used:

- **Accuracy**: The proportion of correctly classified emotions over the total number of samples. Accuracy provides a straightforward measure of the system's overall performance.
- **Precision**: The number of true positive classifications divided by the sum of true positives and false positives. Precision measures the system's ability to avoid false positives (i.e., incorrectly labeling a neutral sample as angry).
- **Recall**: The number of true positive classifications divided by the sum of true positives and false negatives. Recall measures the system's ability to correctly classify all instances of a particular emotion (i.e., correctly identifying all instances of sadness).
- **F1-Score**: The harmonic mean of precision and recall, providing a balanced measure of the system's performance, especially in cases where precision and recall differ significantly.
- **Confusion Matrix**: A detailed matrix that highlights the number of correct and incorrect classifications for each emotion category. The confusion matrix provided insights into which emotions were commonly misclassified and why.

4.6 Results and Analysis

The proposed system demonstrated significant improvements over traditional feature-based approaches in terms of classification accuracy and efficiency. The key results are as follows:

- **Classification Accuracy**: The system achieved an overall accuracy of **92%** on the test set, outperforming other methods such as k-Nearest Neighbors (kNN) and Gaussian Mixture Models (GMM). The combination of MFCC, ZCR, HNR, and TEO features, along with the auto-encoder, allowed the system to capture both spectral and prosodic features critical to emotion recognition.
- **Precision and Recall**: Precision and recall scores for most emotions (happiness, sadness, and anger) were above **90%**, demonstrating the system's ability to correctly identify these emotions with minimal false positives or false negatives. However, emotions like **fear** and **neutral** were slightly more challenging, with precision and recall scores closer to **85%**, likely due to the subtle differences between these emotional expressions in speech.
- **Confusion Matrix**: The confusion matrix revealed that **sadness** and **fear** were occasionally confused, likely due to similarities in prosodic features like low pitch and low energy. Similarly, **neutral** was sometimes misclassified as **happiness** when the speech exhibited slight variations in pitch or energy that resembled happiness.
- **Impact of Auto-Encoder**: The use of the auto-encoder for dimensionality reduction significantly improved classification accuracy and reduced computational complexity. Without the auto-encoder, the system tended to overfit, especially on the training data, leading to lower accuracy on the test set. The reduced 20-dimensional feature vector retained the most relevant information while discarding redundant or noisy data.



4.7 Comparative Analysis

A comparative analysis was conducted to evaluate the performance of the proposed system against other state-of-the-art methods. The system was compared with:

- **CNN-based models**: While deep learning models like CNNs can offer high accuracy, they require extensive computational resources and large amounts of labeled data. Our SVM-based system, while simpler, provided comparable accuracy without the need for large-scale training.
- **Traditional ML classifiers (e.g., kNN, GMM)**: The proposed system outperformed traditional machine learning classifiers, which struggled with the high dimensionality and complexity of the emotional data.

4.8 Discussion of Findings

The results confirm that combining diverse feature extraction techniques with dimensionality reduction and robust classifiers like SVM can significantly enhance the accuracy and efficiency of emotion recognition systems. The system's high accuracy in recognizing emotions like happiness and anger underscores the importance of using both prosodic and spectral features in capturing emotional nuances. The confusion between sadness and fear highlights the need for further research into more sophisticated feature extraction techniques or the incorporation of additional data modalities (e.g., linguistic content or facial expressions) to improve classification accuracy for subtle emotions.

4.9 Limitations

While the system performed well, there are several limitations:

- **Dataset Size**: The RML dataset, though comprehensive, may not represent the full range of emotional diversity found in natural settings. Expanding the dataset to include spontaneous emotional expressions would improve the system's robustness.
- **Real-Time Application**: Although the system shows promise for real-time application, future work will need to focus on optimizing processing time, particularly in environments with limited computational resources

4.10 Future Work

Although the current system achieves strong results, several areas for future development can enhance its performance and adaptability:

1. **Multi-modal Emotion Recognition**

   Future work can integrate other forms of emotional expression, such as facial expressions and physiological signals, alongside speech. This would create a more comprehensive emotion detection system, as emotions are often expressed through multiple modalities simultaneously.

2. **Real-Time Application**

   The system should be optimized for real-time use, reducing processing delays. Faster feature extraction and classification can be achieved through algorithmic improvements or hardware acceleration, enabling its deployment in live scenarios like virtual assistants or mental health monitoring.

3. **Emotion Intensity and Temporal Tracking**

   Rather than focusing only on discrete emotions, future systems should assess the **intensity** of emotions, distinguishing between mild and extreme emotional states. Additionally, tracking how emotions change over time could provide deeper insights during longer interactions.

4. **Multi-lingual and Cross-Cultural Generalization**

   Future research should extend the system to handle multi-lingual datasets and adapt to cultural differences in emotional expression, improving the model’s applicability across diverse populations.

5. **Noise Robustness**

   The system's robustness should be improved to handle noisy environments, such as public places or call centers, ensuring reliable emotion recognition even under challenging conditions.

These advancements will help scale the system for broader, real-world applications, from personal assistants to mental health tools.

5. CONCLUSION

This research presents a robust system for emotion recognition from speech signals, utilizing advanced feature extraction techniques and machine learning models. By combining Mel-Frequency Cepstral Coefficients (MFCC), Zero Crossing Rate (ZCR), Harmonic-to-Noise Ratio (HNR), and Teager Energy Operator (TEO) with dimensionality reduction via an auto-encoder and classification through Support Vector Machines (SVM), the system achieves high accuracy and reliability.

The results demonstrate that the proposed approach effectively captures and classifies emotional states, achieving an overall accuracy of 92% on the Ryerson Multimedia Laboratory (RML) dataset. This performance highlights the system's potential for various applications, including human-computer interaction, mental health monitoring, and customer service.

However, there are several avenues for future improvement. Integrating multi-modal data, such as facial expressions and physiological signals, can provide a more comprehensive understanding of emotions. Enhancing the system's real-time processing capabilities will enable its deployment in interactive applications. Additionally, extending the system to handle emotion intensity, track temporal changes, and generalize across languages and cultures will broaden its applicability and robustness.

Addressing these future directions will advance the field of emotion recognition, making it more adaptable and effective in real-world scenarios. This research paves the way for further innovations, aiming to create more responsive and empathetic systems that better understand and interact with human emotions.

6. REFERENCES

[1] Speech emotion recognition via graph-based representation By Anastasia Pentari, George Kafentzis & Manolis Tsiknaki
[2] Speech Emotion Recognition Using Deep Learning Techniques: A Review
[3] Speech Emotion Recognition -Ashish B. Ingale, D. S. Chaudhari
[4] Speech Emotion Recognition through Hybrid Features and Convolutional Neural Network by Ala Saleh Alluhaidan 1ORCID,Oumaima Saidani 1,*ORCID,Rashid Jahangir 2ORCID,Muhammad Asif Nauman 3 andOmnia Saidani Neffati 4
[5] SPEECH EMOTION RECOGNITION By Shreyansh Puri (191363)
[6] https://images.app.goo.gl/m1y9ix1kR6kwUpR26
[7] A Comprehensive Review of Speech Emotion Recognition Systems TAIBA MAJID WANI 1, TEDDY SURYA GUNAWAN 1,3, (Senior Member, IEEE),SYED ASIF AHMAD QADRI 1, MIRA KARTIWI 2, (Member, IEEE),AND ELIATHAMBY AMBIKAIRAJAH 3, (Senior Member, IEEE)