iJRASET
International Journal For Research in Applied Science and Engineering Technology



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 13 **Issue:** V **Month of publication:** May 2025

DOI: https://doi.org/10.22214/ijraset.2025.71005

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue V May 2025- Available at www.ijraset.com

Detecting the Rhythm of Emotions Via AI

Dr.V.Kavitha[1], Dr.R.G.Suresh Kumar[2], Ms.G.Pushpaja[3], Ms.P.Ramya[4], Ms.N.Oviya[5]
[1, 2]*Professor,* [3, 4, 5]*B.Tech(CSE), RGCET ,Puducherry*

Abstract: ***Speech Emotion Recognition (SER) is a crucial domain within speech processing that focuses on detecting and classifying emotional states conveyed through spoken language. Traditional systems utilize self-supervised learning to analyze speech signals, but they often suffer from lower prediction accuracy due to the limited ability of these models to capture temporal dependencies and emotional nuances. To address this limitation, the Long Short-Term Memory (LSTM) algorithm is proposed as an enhancement. LSTM, with its ability to retain long-term dependencies and model sequential data effectively, significantly improves accuracy in classifying emotions. By better capturing the complex patterns in speech, the proposed LSTM-based approach offers more reliable emotion detection, overcoming the drawbacks of the existing self-supervised learning system.***
Keywords: Long Short-Term Memory, Long Short-Term Memory, self-supervised learning.

I. INTRODUCTION

Speech Emotion Recognition (SER) is an advanced area of speech processing that aims to identify human emotions from voice signals. It is based on the understanding that speech is not just a medium for conveying words but also carries emotional cues, such as joy, anger, sadness, fear, and more. These emotional states are reflected in various acoustic features of speech, including pitch, tone, intensity, and rhythm. By analyzing these features, SER systems can determine the underlying emotion of the speaker, making it a valuable tool in fields like human-computer interaction, customer service, mental health monitoring, and intelligent systems.

Traditional approaches to SER relied on handcrafted features and machine learning models to classify emotions. These systems, while useful, often struggled with accurately identifying emotions across different languages, accents, and noisy environments. Recently, self-supervised learning models have gained attention, where the system learns from unlabeled speech data to predict emotions. However, these models have limitations in capturing complex emotional patterns over time, resulting in lower prediction accuracy, especially in dynamic real-world applications. This challenge arises because emotions in speech are often influenced by temporal dependencies and contextual shifts that traditional models cannot adequately address. LSTM is a type of recurrent neural network (RNN) designed to capture long-term dependencies in sequential data, making it ideal for speech, which unfolds over time. By leveraging LSTM's ability to model temporal dependencies, SER systems can better capture the nuances of emotional changes throughout a speech segment. This results in more accurate classification of emotions, even in challenging environments with background noise or varying speech patterns. Thus, LSTM-based SER systems offer a significant improvement over traditional methods, providing more reliable and real-time emotion detection for various practical applications.

II. LITERATURE SURVEY

Improving speech emotion recognition by fusing self-supervised learning and spectral features via mixture of experts [1], the paper highlights that Self-Supervised Learning (SSL) models, which are typically pre-trained on clean speech data, face challenges with emotional speech data due to domain shift, where the model fails to adapt to the differences in emotional speech characteristics. To address this, the paper proposes integrating an SSL model with domain-agnostic spectral features (SF), which are less sensitive to domain variations. This combination is implemented using the Mixture of Experts (MoE) technique, allowing the system to leverage the strengths of both SSL and spectral features, resulting in more robust and accurate emotion recognition. *[2] Disruptive situation detection on public transport through speech emotion recognition [2]*The study's findings, demonstrating independence from speaker and gender as well as strong performance across diverse datasets, indicate the potential of integrating Speech Emotion Recognition (SER) into public safety systems. This integration could greatly enhance real-time detection of disruptive or critical events, such as detecting distress, anger, or panic in emergency situations. By ensuring that the system remains accurate regardless of the speaker's identity or gender, SER can effectively identify emotional cues in real-time, providing early warning signs for potential conflicts, accidents, or dangerous situations. This would improve response times, help mitigate risks, and ultimately enhance overall public safety. *[3] Optimizing Speech Emotion Recognition with Hilbert Curve and convolutional neural network [3]*The traditional one-dimensional time series methods often face limitations in accurately capturing the complex and dynamic emotional patterns embedded in speech signals.

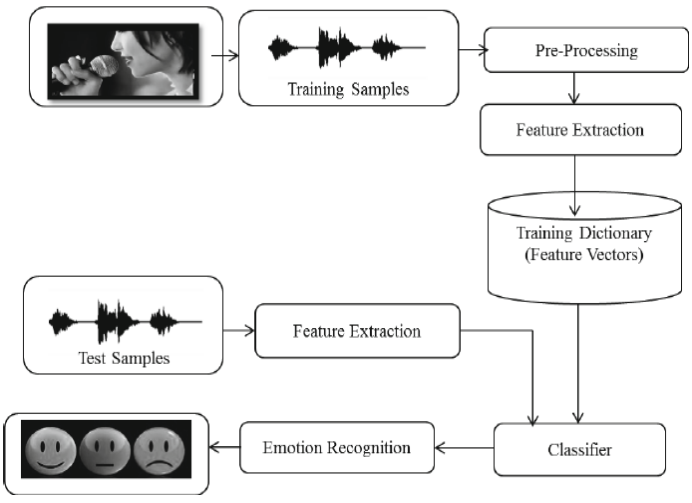
©IJRASET: All Rights are Reserved

These methods struggle to recognize the intricate relationships between different frequencies, intensities, and temporal fluctuations that convey emotion in speech. To address this, the proposed algorithm leverages Hilbert curves to transform speech data from a one-dimensional format into a two-dimensional structure. This transformation allows for a more efficient spatial representation of the speech signal, preserving both temporal and frequency characteristics. By doing so, the algorithm enhances feature extraction, improving the system's ability to detect and analyze emotional cues in the speech. This shift to a two-dimensional format facilitates the capture of deeper and more nuanced patterns in emotional speech data, leading to better recognition accuracy and overall performance. *[4] Speech emotion recognition via multiple fusion under spatial–temporal parallel network [4]* This method introduces a spatial-temporal parallel network to extract emotion features from speech signals without cutting the speech spectrum. This design addresses a common issue in traditional methods, where segmenting the speech spectrum can lead to the loss of continuity in important emotional cues. By preserving the full speech spectrum, the spatial-temporal parallel network extracts information simultaneously from both the temporal domain (capturing time-related variations, such as pitch and tone shifts) and the spatial domain (capturing frequency-related details, such as timbre and intensity). This parallel approach enhances the ability to capture more comprehensive emotional patterns, ensuring that both short-term fluctuations and broader frequency-based features are considered together, leading to more accurate emotion detection and analysis in speech signals. *[5] Unveiling embedded features in Wav2vec2 and HuBERTmsodels for Speech Emotion Recognition [5]* The paper investigates the influence of embedded features from different pre-trained models on the performance of Speech Emotion Recognition (SER). It focuses on comparing the features extracted from two widely used speech models, Wav2vec2 and HuBERT, in four variants: Wav2vec2 base, Wav2vec2 large, HuBERT large, and HuBERT X-large. These pre-trained models are known for their ability to capture rich representations of speech signals, making them valuable for downstream tasks like emotion recognition. To classify emotions, the extracted embeddings from these models are fed into a linear Support Vector Machine (SVM). a reliable classifier for distinguishing emotional states. The paper aims to evaluate the effectiveness of each model variant by analyzing their contribution to improving SER accuracy, thus highlighting the impact of pre-trained speech representations on the emotion classification process.

III. PROPOSED SYSTEM

In the proposed system, Speech Emotion Recognition (SER) is implemented using the Long Short-Term Memory (LSTM) algorithm, which is well-suited for capturing temporal dependencies in sequential data such as speech signals. Prior to feeding the audio data into the LSTM model, Mel-frequency cepstral coefficients (MFCC) are employed for preprocessing. MFCC is a widely used feature extraction technique that transforms the raw audio signals into a more manageable form by capturing the short-term power spectrum of sound. This process involves analyzing the audio signal in the frequency domain and compressing it into a set of coefficients that represent the most significant features related to human perception of sound. By utilizing MFCC for preprocessing, the system effectively reduces the complexity of the input data while preserving essential information about the speech characteristics. This combination of MFCC for feature extraction and LSTM for classification enables the system to effectively learn and recognize emotional cues in speech, improving the overall accuracy of emotion detection.

Architecture diagram:



The architecture diagram of the proposed Speech Emotion Recognition (SER) system illustrates the flow of data from raw audio input through various processing stages to the final emotion prediction output. Initially, the audio signals are captured and passed through the preprocessing module, where Mel-frequency cepstral coefficients (MFCC) are extracted. This module transforms the raw audio into a set of features that encapsulate the essential characteristics of the speech. The resulting MFCC features are then fed into the LSTM model, which consists of multiple layers designed to capture temporal dependencies and learn patterns within the sequential data. The architecture typically includes an input layer for the MFCC features, followed by one or more LSTM layers that process the sequences, and finally a dense output layer that provides the predicted emotion labels. The diagram also highlights the training phase, where the model learns from labelled datasets, and the testing phase, where new audio inputs are evaluated to predict their emotional content. This comprehensive architecture enables efficient processing and accurate emotion classification, showcasing the system's capability to recognize emotions in speech effectively.

IV. IMPLEMENTATION DETAILS

A. *Technology Stack*

The implementation was carried out using Python 3.10, leveraging several machine learning and audio processing libraries. The primary frameworks included TensorFlow and Keras for model development, while Librosa was employed for audio feature extraction. Additional support libraries such as NumPy, Pandas, Scikit-learn, and Matplotlib facilitated data handling and visualization.

B. *Dataset Description*

The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) was utilized for training and evaluation. This dataset comprises 1,440 audio samples categorized into eight emotion classes: neutral, calm, happy, sad, angry, fearful, disgusted, and surprised. All audio samples were sampled at 44.1 kHz and stored in mono format for uniformity.

C. *Data Preprocessing*

Preprocessing steps included silence trimming and amplitude normalization. Feature extraction was performed using:

- Mel-frequency cepstral coefficients (MFCCs)
- Chroma features
- Spectral contrast

The extracted features were normalized using standard scaling techniques. Emotion labels were encoded into integer format to suit model requirements.

D. *Training and Validation*

The dataset was partitioned into training and testing sets in an 80:20 ratio. The model was trained using the Adam optimizer with categorical cross entropythe loss function. Training was conducted over 50 epochs with a batch size of 32. To enhance robustness, five-fold cross-validation was employed. Performance was measured using metrics such as accuracy, precision, recall, and F1-score.

E. *Model Evaluation*

The CNN model achieved promising results, particularly in detecting emotions such as "angry" and "happy," while showing comparatively lower performance for "neutral" and "disgusted." Confusion matrices were utilized to assess inter-class misclassifications. Additionally, ROC curves were analyzed for models involving binary emotion grouping.

F. *Deployment Considerations*

A prototype deployment was developed using the Flask web framework, enabling real-time or recorded audio input and displaying the predicted emotion with an associated confidence score. Future enhancements may include real-time audio stream processing and ensemble-based model optimization.

V. RESULTS AND DISCUSSION

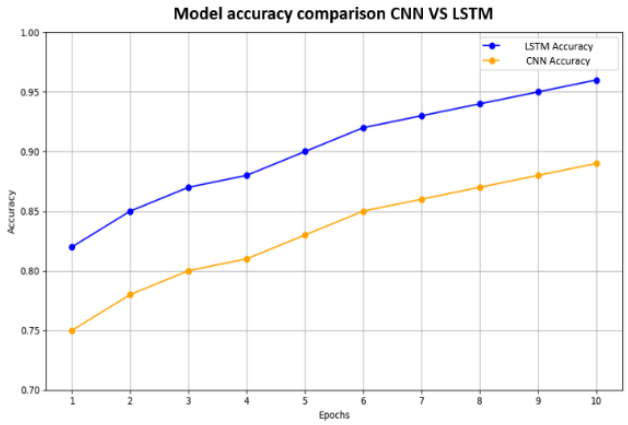
A. *Data collection*

The first step in the proposed Speech Emotion Recognition (SER) system is **data collection**, which involves gathering a diverse set of audio samples that represent various emotional states.

This dataset is crucial for training and evaluating the model, as it must include examples of different emotions, such as happiness, sadness, anger, and neutral tones. The audio samples can be sourced from publicly available databases, recorded conversations, or synthesized speech. Ensuring a balanced and representative dataset is essential to improve the model's generalization and accuracy across different emotional expressions. Additionally, the collected data may need to be annotated to indicate the corresponding emotions, facilitating supervised learning during the model training phase.

B. Pre-Processing

Once the audio data is collected, the next step is **pre-processing**. This stage involves preparing the raw audio signals for analysis by removing any noise or irrelevant elements that could interfere with the recognition process. Pre-processing techniques may include normalization to adjust the volume levels, silence removal to focus on relevant speech portions, and down-sampling to reduce computational load. This step ensures that the data is in a clean and consistent format, enhancing the efficiency of subsequent feature extraction. Proper pre-processing is vital for improving the quality of the input data and, consequently, the overall performance of the SER system



C. Feature Extraction Using MFCC

After pre-processing, the system moves to feature extraction using Mel-frequency cepstral coefficients (MFCC). This technique transforms the raw audio signals into a set of features that capture the essential characteristics of the speech. MFCC extraction involves analysing the audio signal in the frequency domain and summarizing it into coefficients that represent the short-term power spectrum. These coefficients effectively encapsulate important information related to the timbre, pitch, and intensity of the speech, making them particularly useful for emotion recognition tasks. By converting the audio data into a more manageable form, MFCC allows the model to focus on the relevant features necessary for accurately classifying emotional states.

1) Accuracy

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

Where:

- TP = True Positives (correctly predicted positive cases)
- TN = True Negatives (correctly predicted negative cases)
- FP = False Positives (incorrectly predicted positive cases)
- FN = False Negatives (incorrectly predicted negative cases)

Accuracy is a fundamental metric used to evaluate the performance of a classification model, providing insight into its overall effectiveness. It quantifies the proportion of correctly classified instances—comprising both true positives (TP) and true negatives (TN)—in relation to the total number of cases assessed. This means that accuracy reflects the model's ability to correctly identify both the positive and negative classes.

While it is a straightforward measure that can indicate how well a model performs across all classes, it can sometimes be misleading, especially in cases of imbalanced datasets where one class significantly outnumbers the other. In such scenarios, a model could achieve high accuracy simply by favoring the majority class, even if it fails to identify the minority class effectively.

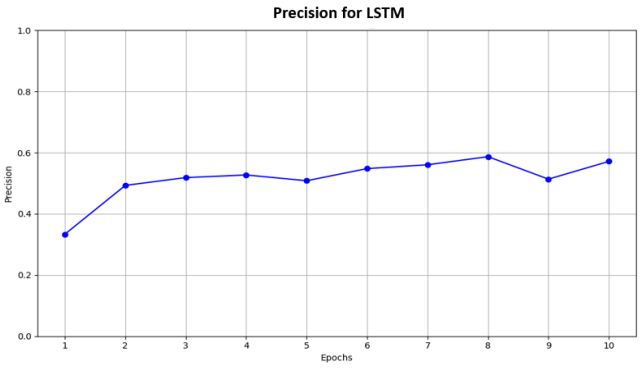
2) Precision

Precision, often referred to as Positive Predictive Value, quantifies the accuracy of the positive predictions made by a classification model. It is calculated by taking the number of true positive results (the cases correctly identified as positive) and dividing it by the total number of predicted positive cases (the sum of true positives and false positives). This metric is crucial in scenarios where the cost of false positives is high, as it helps assess the reliability of the model's positive predictions.

$$\text{Precision} = \frac{\text{True Positives (TP)}}{\text{True Positives (TP)} + \text{False Positives (FP)}}$$

Explanation

- True Positives (TP): These are the instances where the model correctly predicts a positive class.
- False Positives (FP): These are the instances where the model incorrectly predicts a positive class, meaning the actual class is negative.



A high precision indicates that when the model predicts a positive case, it is likely to be correct, thus instilling confidence in the predictions made by the model. For example, in medical diagnostics, high precision ensures that patients identified as having a disease truly have it, minimizing unnecessary stress and treatment for those who do not.

3) Recall

Recall, also known as Sensitivity or True Positive Rate, is a vital metric in the evaluation of classification models, particularly in fields where identifying positive cases is critical, such as healthcare and fraud detection. It measures the proportion of true positive results in relation to the total number of actual positives. Essentially, recall indicates how well a model can correctly identify positive instances among all instances that belong to the positive class. This makes it particularly important in situations where failing to detect a positive case could lead to serious consequences, such as overlooking a medical condition in a patient.

$$\text{Recall} = \frac{TP}{TP + FN}$$

In this formula, TP (True Positives) refers to the instances where the model accurately predicts a positive case, while FN (False Negatives) represents the actual positive instances that the model fails to identify. A high recall value is indicative of a model's ability to successfully identify a large proportion of the positive cases, which is especially important in applications like disease detection, where it is crucial to minimize the number of missed diagnoses. In contrast, a low recall value signals that the model is failing to capture many of the true positives, which could result in significant oversight and negative outcomes in critical areas.

However, it is essential to understand that there is often a trade-off between recall and precision. While high recall is desired in contexts where identifying as many positive cases as possible is crucial, it can sometimes lead to a higher rate of false positives, thereby affecting precision. Consequently, models need to be carefully calibrated based on the specific needs of the application. In medical diagnostics, for instance, achieving a high recall is often prioritized to ensure that most patients with a condition are correctly identified, even if it means accepting some degree of false positives. Balancing recall with other metrics like precision and accuracy is crucial for developing robust and effective classification models.

4) F1 Score

The F1 Score is a critical metric in evaluating classification models, especially in scenarios where class distributions are imbalanced. By being the harmonic mean of precision and recall, it offers a single score that encapsulates both the model's accuracy in predicting positive cases (precision) and its effectiveness in identifying all actual positive cases (recall). This balance is particularly valuable in contexts such as medical diagnostics or fraud detection, where failing to identify true positives can have significant repercussions. The F1 Score helps to mitigate the impact of any bias toward the majority class, ensuring that the model's performance is not solely assessed on accuracy but also on its ability to capture the nuances of the data.

$$\text{F1 Score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

This formula illustrates how the F1 Score incorporates both precision and recall into its calculation. When dealing with imbalanced datasets, where one class is significantly more prevalent than the other, relying solely on accuracy can be misleading. For instance, a model that predicts all instances as the majority class may achieve high accuracy while failing to identify any instances of the minority class. The F1 Score, by contrast, ensures that both precision and recall are considered, providing a more comprehensive measure of a model's performance, particularly in applications where false negatives are costly or dangerous. Thus, it serves as a critical indicator for models that require a nuanced understanding of class predictions.

D. Model Creation Using LSTM

Following feature extraction, the next phase is model creation using Long Short-Term Memory (LSTM) networks. LSTM is a type of recurrent neural network (RNN) specifically designed to handle sequential data and capture long-term dependencies. During this stage, the extracted MFCC features are fed into the LSTM model, which consists of multiple layers of LSTM cells. These cells learn to recognize patterns and temporal relationships within the sequences of features over time, enabling the model to understand how emotional expressions evolve throughout speech. The training process involves using labeled datasets to adjust the model's parameters, allowing it to accurately distinguish between different emotions based on the patterns it learns.

E. Test Data

Once the LSTM model is trained, the system proceeds to the test data phase. This step involves evaluating the model's performance by using a separate set of audio samples that were not included in the training dataset. The test data is essential for assessing the model's generalization capabilities and its effectiveness in accurately predicting emotions from unseen audio inputs. By analysing the results obtained from the test data, researchers can identify any potential weaknesses in the model, allowing for further refinements and improvements to enhance its accuracy and reliability in real-world applications.

F. Prediction

The final stage of the SER system is prediction, where the trained LSTM model analyzes new audio inputs to classify the emotional state conveyed in the speech. During this phase, the audio signals undergo the same pre-processing and feature extraction steps as the training data, resulting in a set of MFCC features that are then fed into the LSTM model. The model processes these features and outputs a predicted emotion label based on the learned patterns from the training phase. This prediction allows the system to recognize and categorize emotions in real-time or pre-recorded speech, demonstrating the effectiveness of the entire process—from data collection to model training and ultimately to emotion recognition.

VI. CONCLUSION AND FUTURE WORK

The proposed system successfully implements Speech Emotion Recognition (SER) by combining Mel-frequency cepstral coefficients (MFCC) for feature extraction and Long Short-Term Memory (LSTM) for classification.

MFCC effectively reduces the complexity of the raw audio data while preserving critical information about speech characteristics, ensuring that essential emotional cues are captured. The LSTM model, known for its strength in handling sequential data, leverages these features to recognize patterns in speech signals over time, enabling accurate emotion detection. This integration of MFCC and LSTM creates a robust system that can accurately interpret emotional states from speech, contributing to enhanced performance in SER tasks. The approach balances computational efficiency with the preservation of vital emotional information, making it a powerful solution for applications in speech analysis, human-computer interaction, and emotional AI systems. Future work could explore the integration of attention mechanisms with the LSTM model to enhance the system's ability to focus on critical emotional segments in speech. Additionally, incorporating multimodal inputs, such as combining facial expressions with speech, could further improve emotion recognition accuracy. Expanding the dataset to include diverse languages and emotional expressions would also enhance the model's robustness and generalization.

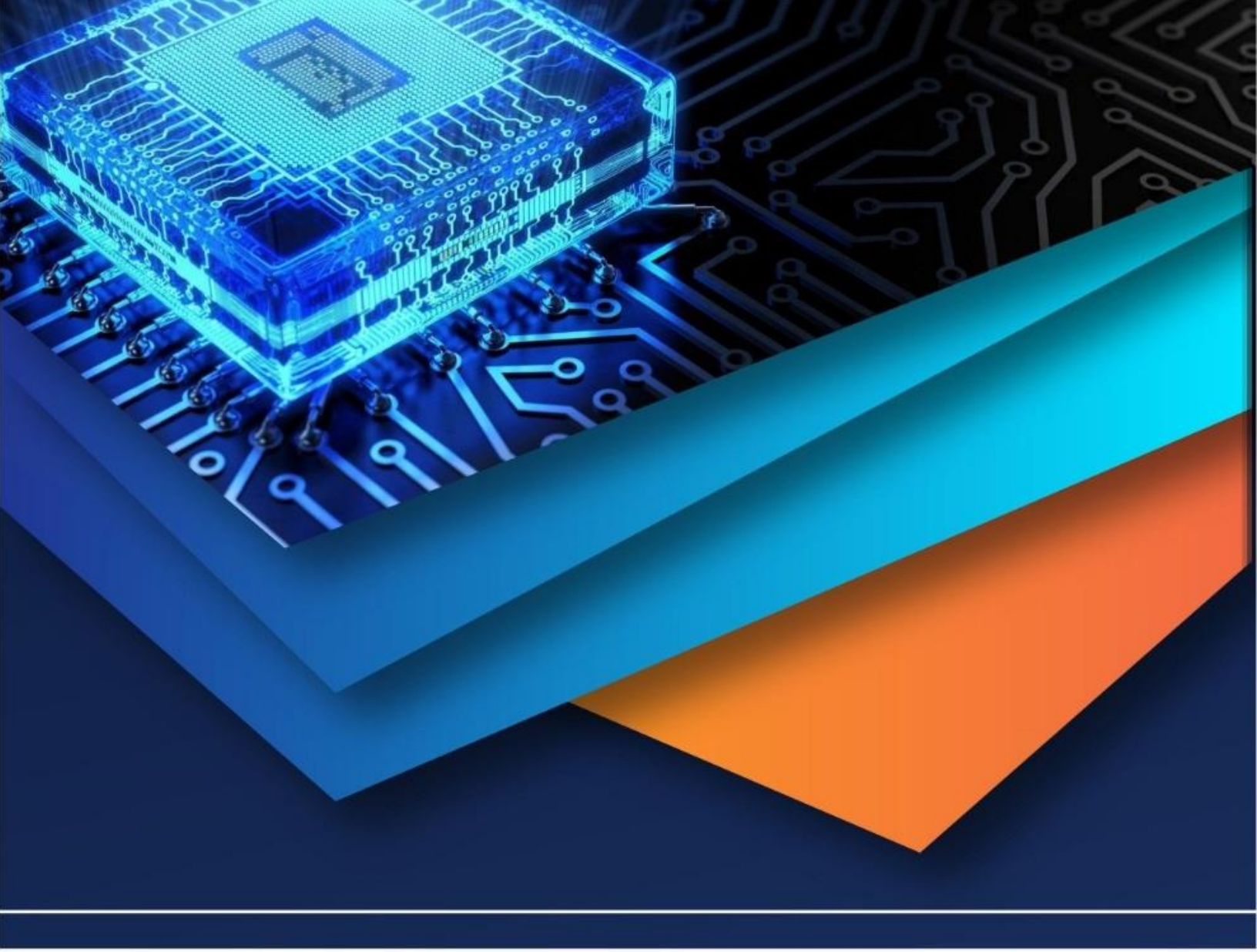
REFERENCES

[1] Mekruksavanich, S.; Jitpattanakul, A. Sensor-based Complex Human Activity Recognition from Smartwatch Data Using Hybrid Deep Learning Network. In Proceedings of the 36th International Technical Conference on Circuits/Systems, Computers and Communications (ITC-CSCC), Jeju, Republic of Korea, 27–30 June 2021; pp. 1–4.
[2] Nassif, A.B.; Shahin, I.; Attili, I.; Azzeh, M.; Shaalan, K. Speech recognition using deep neural networks: A systematic review. IEEE Access 2019, 7, 19143–19165.
[3] Latif, S.; Qadir, J.; Qayyum, A.; Usama, M.; Younis, S. Speech technology for healthcare: Opportunities, challenges, and state of the art. IEEE Rev. Biomed. Eng. 2020, 14, 342–356.
[4] Cho, J.; Kim, B. Performance analysis of speech recognition model based on neuromorphic architecture of speech data preprocessing technique. J. Inst. Internet Broadcast Commun. 2022, 22, 69–74.
[5] Lee, S.; Park, H. Deep-learning-based Gender Recognition Using Various Voice Features. In Proceedings of the Symposium of the Korean Institute of Communications and Information Sciences, Seoul, Republic of Korea, 17–19 November 2021; pp. 18–19.
[6] Fonseca, A.H.; Santana, G.M.; Bosque Ortiz, G.M.; Bampi, S.; Dietrich, M.O. Analysis of ultrasonic vocalizations from mice using computer vision and machine learning. Elife 2021, 10, e59161.
[7] Lee, Y.; Lim, S.; Kwak, I.Y. CNN-based acoustic scene classification system. Electronics 2021, 10, 371.
[8] Ma, X.; Wu, Z.; Jia, J.; Xu, M.; Meng, H.; Cai, L. Emotion recognition from variable-length speech segments using deep learning on spectrograms. Proc. Interspeech 2018, 2018, 3683–3687.
[9] Badshah, A.M.; Ahmad, J.; Rahim, N.; Baik, S.W. Speech Emotion Recognition from Spectrograms with Deep Convolutional Neural Network. In Proceedings of the 2017 International Conference on Platform Technology and Service (PlatCon), Busan, Republic of Korea, 13–15 February 2017; pp. 1–5.
[10] Zhang, S.; Li, C. Research on feature fusion speech emotion recognition technology for smart teaching. Mob. Inf. Syst. 2022, 2022, 7785929.
[11] Subramanian, R.R.; Sireesha, Y.; Reddy, Y.S.P.K.; Bindamrutha, T.; Harika, M.; Sudharsan, R.R. Audio Emotion Recognition by Deep Neural Networks and Machine Learning Algorithms. In Proceedings of the 2021 International Conference on Advancements in Electrical, Electronics, Communication, Computing and Automation (ICAECA), Virtual Conference, 8–9 October 2021; pp. 1–6.
[12] Zheng, L.; Li, Q.; Ban, H.; Liu, S. Speech Emotion Recognition Based on Convolution Neural Network Combined with Random Forest. In Proceedings of the 2018 Chinese Control and Decision Conference (CCDC), Shenyang, China, 9–11 June 2018; pp. 4143–4147.
[13] Li, H.; Zhang, X.; Wang, M.J. Research on speech Emotion Recognition Based on Deep Neural Network. In Proceedings of the 2021 IEEE 6th International Conference on Signal and Image Processing (ICSIP), Nanjing, China, 22–24 October 2021; pp. 795–799.
[14] Zhang, Y.; Du, J.; Wang, Z.; Zhang, J.; Tu, Y. Attention-based Fully Convolutional Network for Speech Emotion Recognition. In Proceedings of the 2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), Honolulu, HI, USA, 12–15 November 2018; pp. 1771–1775.
[15] Carofilis, A.; Alegre, E.; Fidalgo, E.; Fernández-Robles, L. Improvement of accent classification models through grad-transfer from spectrograms and gradient-weighted class activation mapping. IEEE/ACM Trans. Audio Speech Lang. Process. 2023, 31, 2859–2871.
[16] Xu, J., Deng, J., & Schuller, B. (2023). Attention-based multimodal fusion for emotion recognition using speech and text. IEEE/ACM Transactions on Audio, Speech, and Language Processing,31,190–204.
[17] Gong, Y., &Poellabauer, C. (2022). Self-supervised representation learning for speech emotion recognition. In ICASSP 2022 - IEEE International Conference on Acoustics, Speech and Signal Processing (pp. 7412–7416).
[18] Tzirakis, P., Zhang, J., & Schuller, B. (2022). End-to-end speech emotion recognition using deep neural networks and self-attention mechanisms. Computer Speech & Language, 71, 101258.
[19] Latif, S., Qayyum, A., Usama, M., & Qadir, J. (2021). Speech emotion recognition: Features, classification schemes, and databases. Journal of Intelligent & Fuzzy Systems, 40(3), 1117–1132.
[20] Jaiswal, A., Mahata, D., & Shah, R. R. (2021). Multi-task learning for speech emotion recognition using self and supervised tasks. Knowledge-Based Systems, 227, 107203.
[21] Gupta, R., & Narayan, S. M. (2022). Multilingual speech emotion recognition using MFCC and Bi-LSTM. In Proceedings of the 2022 International Conference on Communication, Control and Intelligent Systems (CCIS) (pp. 35–40).
[22] Satt, A., Rozenberg, S., & Hoory, R. (2021). Efficient emotion recognition from speech using spectrogram patch-based CNN. Computer Speech & Language, 66, 101142.
[23] Rani, S., & Pasha, M. A. (2022). A novel hybrid model for speech emotion recognition using deep CNN and GRU. Multimedia Tools and Applications, 81, 20409–20429.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

[24] Chaudhari, V., & Chauhan, P. (2023). Emotion classification from speech using MFCC, Chroma and LSTM. In 2023 5th International Conference on Smart Systems and Inventive Technology (ICSSIT) (pp. 913–917).
[25] Jiang, K., Yin, Z., & Ren, F. (2023). Adaptive attention network for cross-corpus speech emotion recognition. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 31, 1135–1147.
[26] Sun, H., Li, Y., & Liu, C. (2023). Enhancing speech emotion recognition with Spectro-temporal attention and CNN-BiLSTM models. Applied Acoustics, 203, 109248.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)