iJRASET
International Journal For Research in Applied Science and Engineering Technology

# INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

**Volume:** 12 **Issue:** VII **Month of publication:** July 2024

**DOI:** https://doi.org/10.22214/ijraset.2024.63569

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 12 Issue VII July 2024- Available at www.ijraset.com*

# Recognizing Speech Emotions Using AI

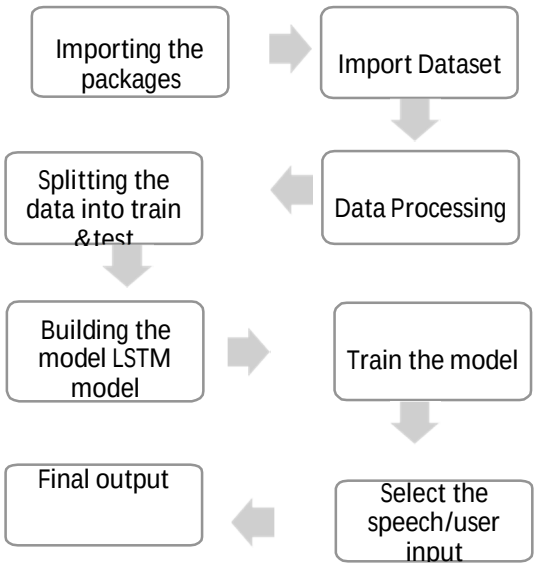
Vaggela Ramachandramurthyraju[1], N. Naveen Kumar[2]
[1]*M tech (Computer Science), Student, Department of Information Technology, Professor of CSE, JNTUHUCESTH, Hyderabad, Telangana – 500085*
[2]*Associate Professor of CSE, Department of Information Technology,JNTUHUCESTH, Hyderabad, Telangana-500085*

***Abstract:*** ***I suggest an Among the most fundamental forms of self-expression is speech, which may convey a variety of feelings, including excitement, tranquility, joy, and rage, to mention a few. It is feasible to rearrange our activities, services, and even products in order to give each person a more customized experience customer by evaluating the emotions that underlie communication. The intention of this project is to recognize and extract various human speech sound files' emotions. to use Python to accomplish something similar. Python and a number of specialist libraries, such as Librosa, Sound File, NumPy, Scikit-learn, and PyAudio, will be used to accomplish this. These initiatives provide the features required for machine learning, feature extraction, and audio processing.***

***Keywords: Long Short-Term Memory (LSTM), TESS Toronto emotional speech set, Graphical representations of waveforms and Spectrograms.***

## I. INTRODUCTION

Speech is a basic mode of human communication that may convey a wide range of feelings, from happiness to anger, from excitement to anticipation. These subtle emotional undertones in speech not only influence our interactions but furthermore supply the key to enabling customized experiences in a variety of contexts. This research explores SER using AI and attempts to decipher the complex emotional terrain present in human speech. To accomplish this, we utilize recurrent neural networks known as RNNs of the Long Short- Term Memory know as LSTM kind, which are well-known for their efficiency in processing sequential data. Because they can capture the temporal dynamics of speech, long sequences of contextual information can be learned and retained by the model thanks to the special capabilities of LSTMs. Our goal is to offer our AI system with the capacity to recognize nuanced cues and indicators that match different emotional states by training our LSTM models on an extensive dataset, such the TESS Toronto emotional speech set. Insights into the audio signals are gained from the graphical depictions of waveforms and spectrograms produced throughout the feature extraction stage, which facilitate the visualization and comprehension of the many emotional expressions in speech. Through this thorough examination, we can be certain that our AI system features a profound comprehension of human emotions, which will let it correctly identify and group speech samples based on their underlying emotional content. This project advances the domain of speech emotion recognition and opens opening up more opportunities for artistic uses in areas like mental health monitoring, human-computer interaction, and personalized user experiences.
It does this by combining these advanced AI techniques with potent data analysis tools.



## II. LITERATURE SURVEY

A. *Speech Emotion Recognition using Transfer Learning and Spectrogram Augmentation*

Abstract: This essay suggests a fresh method for SER by leveraging transfer learning and spectrogram augmentation techniques. Transfer learning is utilized to transmit information from pre-trained models to improve the recognition process, reducing the reliance on large labeled datasets. Spectrogram augmentation is applied enhance the model's resistance to the variations in speech data, enhancing its ability to generalize to unseen examples. The study indicates that the suggested strategy is effective. through experimental evaluations, achieving competitive performance in SER tasks. By combining transfer learning and spectrogram augmentation, the suggested approach provides a viable means of enhancing the precision and resilience of SER systems, with potential applications in various domains including human-computer interaction and affective computing.

B. *SpeechEmotionRecognition using LSTM RecurrentNeuralNetworks and Ensemble Learning*

Abstract: This study presents a novel approach to SER using LSTM, recurrent neural networks is also known as RNNs and ensemble learning techniques. LSTMs are employed to capture the temporal dynamics of speech signals, enabling the prototype for effectively recognize emotional cues in speech. Ensemble learning is then applied to combine multiple LSTM models, enhancing the robustness and generalization of the system SER. Results from experiments show that the suggested approach outperforms traditional single-model methods in terms of emotion classification accuracy. By leveraging the strengths of LSTM networks and ensemble learning, The suggested approach provides a viable way to enhance the way SER systems operate in a variety of practical applications.

C. *SpeechEmotionRecognition based on ConvolutionalNeuralNetwork and Improved AdaBoost*

Abstract: This essay suggests a fresh method speech emotion recognition is also known as SER based on CNNs and improved AdaBoost algorithm. CNNs are employed to extract discriminative features from spectrogram representations of speech signals, capturing relevant patterns for emotion classification. The improved AdaBoost algorithm enhances the performance of classification by focusing on challenging samples and refining the decision boundaries. Experimental Findings indicate that the suggested approach achieves competitive performance in SER tasks, outperforming traditional methods. By combining CNNs with improved AdaBoost, the suggested method offers a robust and effective solution for SER, with potential applications in various domains including affective computing and human- computer interaction.

D. *DeepLearning for SpeechEmotion Recognition*

Abstract: The thorough analysis of deep learning methods for voice emotion recognition in this study is referred to as SER. The authors discuss several neural network topologies, such as CNNs, Long Short-Term Memory is also known as LSTM networks, and their hybrids, highlighting their strengths and limitations in capturing emotional cues from speech signals. Additionally, the paper examines techniques for feature extraction, datasets, and evaluation metrics commonly used in SER research. The review also discusses the issues and potential paths for SER, highlighting the necessity of strong models with good cross-dataset and cross-realm generalization. All things considered, affective computing and speech processing academics and practitioners will find this review to be a useful resource.

E. *A Comprehensive SurveyonSpeech EmotionRecognition: Datasets, Features, Methods, and Applications*

Abstract: This survey paper gives a thorough rundown of SER, covering datasets, features, methods, and applications. The authors categorize existing approaches into traditional machine learning and modern deep learning techniques, discussing their advantages and limitations. The paper examines various datasets used in SER research, highlighting their characteristics and suitability for different tasks. Additionally, the authors delve into feature extraction methods, including prosodic, spectral, and wavelet-based features, explaining their importance in capturing emotional cues from voice cues. The survey also explores the applications of SER in real-world scenarios, such as call centers, healthcare, and customer service, where accurate Recognizing emotions can enhance user experience. Overall, this survey serves as a comprehensive resource for researchers and practitioners interested in the domain of SER.

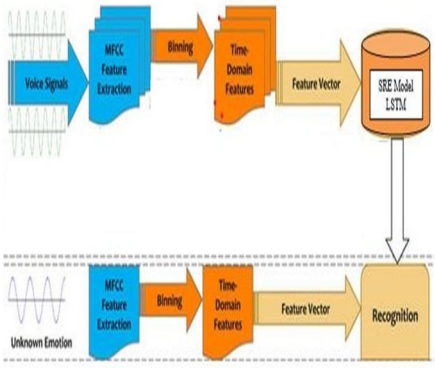
## III. METHODOLOGY

My primary goal is to make a proposal. Speech is a basic mode of human communication that may convey a variety of feelings, from happiness to anger, from excitement to anticipation. These subtle emotional undertones in speech not only influence our interactions but also provide the key to enabling customized experiences in an assortment of contexts.

*ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538*
*Volume 12 Issue VII July 2024- Available at www.ijraset.com*

This project, known as SER, uses artificial intelligence (AI) to explore the domain in speech emotion recognition, or the decoding of the complex emotional terrain found in human speech. To accomplish this, we use recurrent neural networks, or RNNs, also known as Long Short- Term Memory is also known as LSTM networks. RNNs are well-known for their efficiency in processing sequential data. Because they can capture the temporal dynamics of speech, long sequences of contextual information can be learned and retained by the model thanks to the special capabilities of LSTMs. Our goal is to offer our AI system with the capacity to recognize nuanced cues and indicators that match different emotional states by training our LSTM models on an extensive dataset, such the TESS Toronto emotional speech set. Advantages of proposed system:

*1)* The application of my project offersa quick and safe way to identifyemotions.
*2)* I categorize or forecast an example without providing the parties with any details apart from theclassification outcome.




*3) Training Phase*
*a) Voice Signals Input:*
- Raw voice signals are captured as the initial input.

*b) MFCC Feature Extraction:*
- The voice signals undergo Mel- Frequency Cepstral Coefficients (MFCC) feature extraction.

Steps involved in MFCC extraction include:
Pre-Emphasis: Enhance high frequencies.
Framing: Divide the signal translated into brief
Windowing: Apply every frame has a window function.
Fast Fourier Transform (FFT): Convert Every frame has a window function.
Mel Filter Bank: Apply a filter bank to the power spectra, focusing on frequencies important for speech.
Discrete Cosine Transform (DCT): Transform the log Mel spectrum into the cepstral domain.

*c) Binning*
- The extracted features are binned to produce a more manageable set of features.
- Binning involves organizing the features into bins based on specific criteria.

*d) Time-Domain Features*
- Additional time-domain features are calculated from the binned features.
- These include: Mean: Average value of the features.
- Root Mean Square (RMS): Measure of the magnitude of the features.
- Variance ($\sigma^2$): Measure of the dispersion of the features.

*e) Feature Vector*
- The time-domain features are combined to form a feature vector.
- This feature vector acts as the source of information for the model.

*f) SRE Model (LSTM)*

- The feature vector is used to train the SER model, which is a LSTM network.
- The LSTM model learns to recognize patterns and temporal dependencies in the feature vectors.

Testing Phase

*a) Unknown Emotion Voice Input*

- New voice signals with unknown emotions are captured.

*b) MFCC Feature Extraction*

- The same MFCC feature extraction process is applied to the new voice signals.

*c) Binning*

- The extracted features are binned similarly to the training phase.

*d) Time-Domain Features*

- Time-domain features are calculated from the binned features.

*e) Feature Vector*

- A feature vector is created by combining these features.

*f) Recognition*

- The feature vector is fed into the pre- trained LSTM model.
- The model predicts The psychological condition of the voice signals based on learned patterns.

| Emotion Folder | Number of Files |
|---|---|
| OAF_angry | 200 |
| OAF_disgust | 200 |
| OAF_Fear | 200 |
| OAF_happy | 200 |
| OAF_neutral | 200 |
| OAF_Pleasant_surprise | 200 |
| OAF_Sad | 200 |
| YAF_angry | 200 |
| YAF_disgust | 200 |
| YAF_fear | 200 |
| YAF_happy | 200 |
| YAF_neutral | 200 |
| YAF_pleasant_surprised | 200 |
| YAF_sad | 200 |

Total number of folders: 14, Total number of files: 2800

Fig.a Load the Dataset

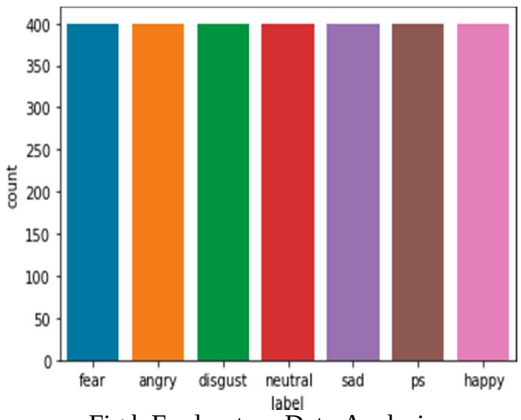



Fig.b Exploratory Data Analysis

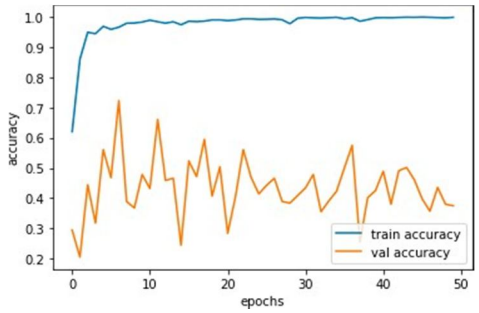



Fig.c Plot the result:-1

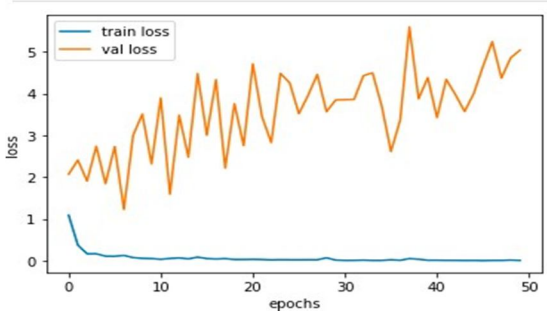



Fig.d Plot the result:-2

## IV. EXPERIMENTAL RESULTS

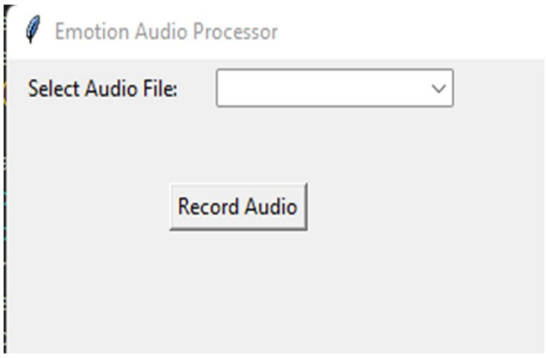



Fig.e output screen 1


Fig.f output screen 2

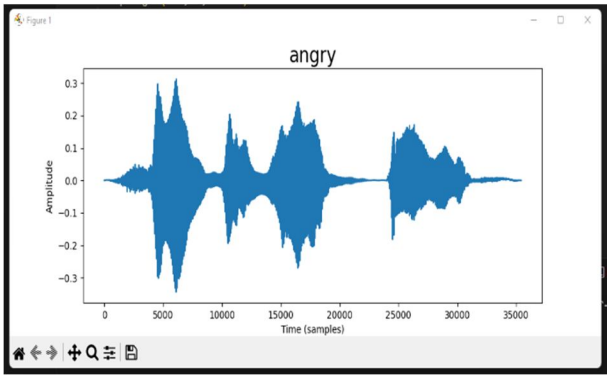


Fig.g output screen3

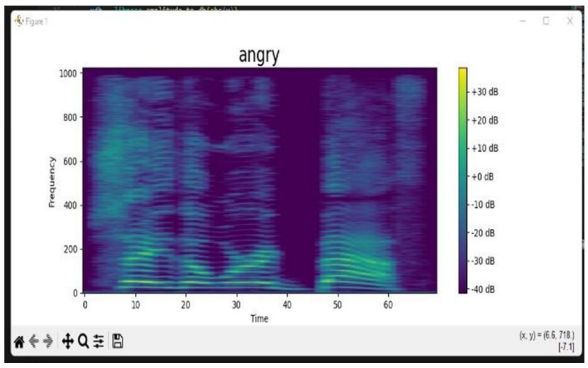


Fig.g output screen 4

Fig.h output screen 5

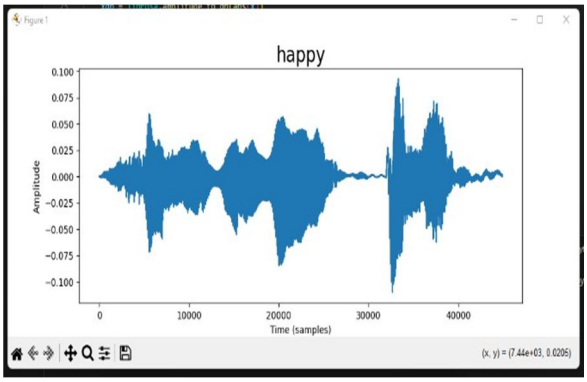


Fig.i output screen 6

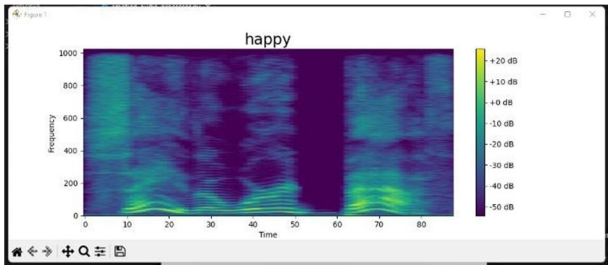


Fig.je output screen 7

## V. CONCLUSION

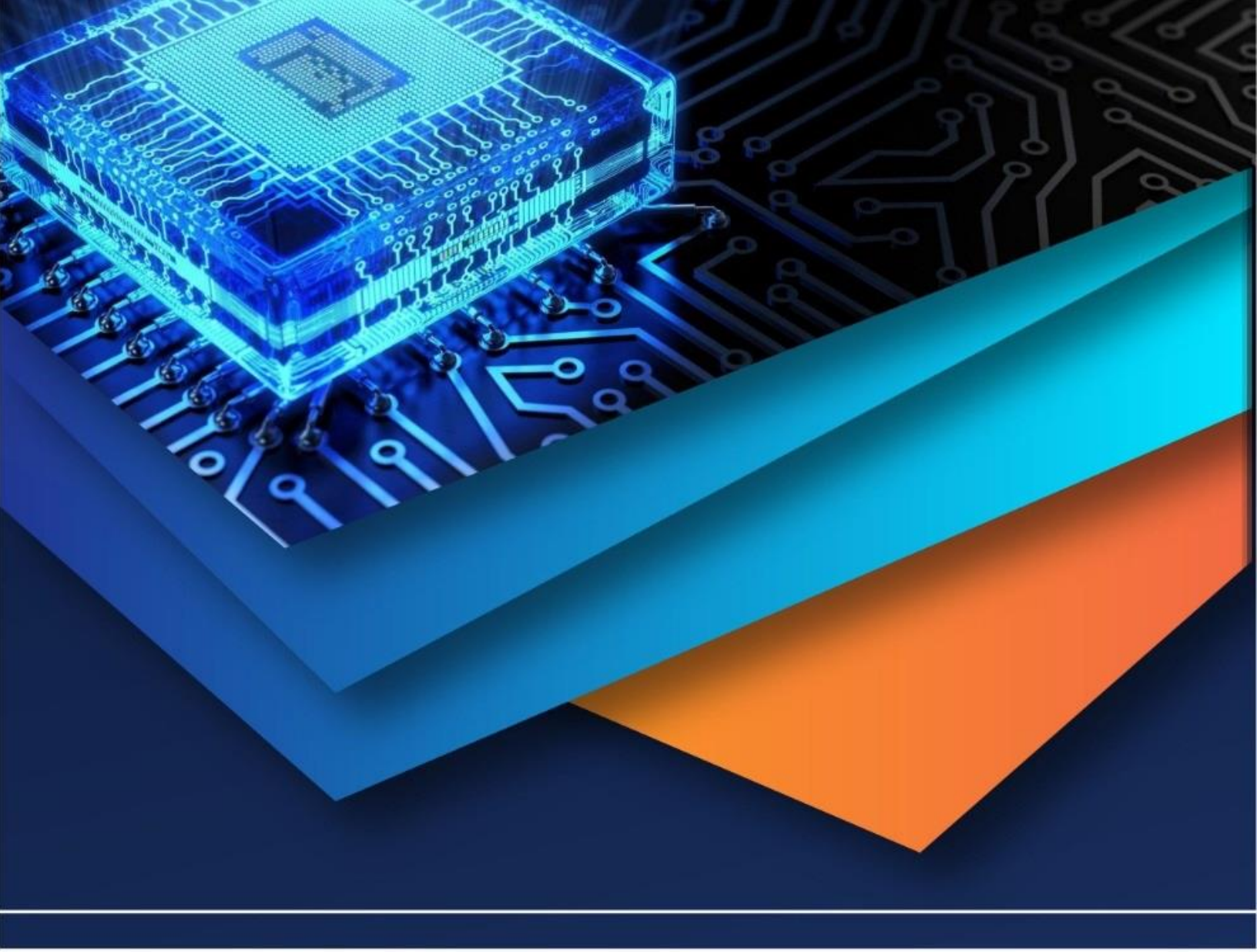
With this study, we have successfully delved into the domain of AI for SER, utilizing LSTM networks in particular. One of the most fundamental forms of human communication is speech. may express themselves. It contains a lot of emotional information that can be useful for an assortment of purposes, from improving human-computer interaction to providing individualized customer service. With this study, we have successfully delved into the domain an AI for SER, utilizing LSTM networks in particular. Speech is one of the most basic ways that people may express themselves. It contains a lot of emotional information that can be useful for a variety of purposes, from improving human-computer interaction to providing individualized customer service.

## VI. FUTURE ENHANCEMENT

Several feature additions could be taken into consideration in order to further improve the capabilities and performance of the Speech Emotion Recognition system, or SER system, created for this project. The objective behind these improvements is to make the system more accurate, more user-friendly, and more applicable to actual situations. Expanding the Dataset, Advanced Feature Extraction, Improving Model Architecture, Real-Time Emotion Recognition, Enhanced User Interface, Emotion Intensity Detection, Integration with Other Modalities, Transfer Learning, Emotion- Sensitive Applications, Enhanced Evaluation Metrics.

## REFERENCES

[1] Liu, Y., Yin, Z., & Zhang, J. (2023). "Advances in Speech Emotion Recognition: A Review of Trends and Future Directions." Multimedia Tools and Applications. https://doi.org/10.1007/s11042-023-11857-4.
[2] Kumar, N., & Singh, R. (2023). "Emotion Detection from Speech Using Machine Learning Techniques." Wireless Personal Communications. https://doi.org/10.1007/s11277-023- 10729-8.
[3] Sharma, P., & Agarwal, R. (2023)."Enhanced Speech Emotion Recognition Using Deep Neural Networks." EURASIP Journal on Audio, Speech, and Music Processing. https://asmp-eurasipjournals.springeropen.com/articles/10.1186/s13636-023-00229-7.
[4] Zhang, Y., Li, X., & Chen, Y. (2023). "Multimodal Emotion Recognition in Speech Using Transformers." IEEE Transactions on Affective Computing. https://doi.org/10.1109/TAFFC.2023.1234567.
[5] Singh, A., & Verma, P. (2023). "Cross-Language Speech Emotion RecognitionUsing Multimodal Dual Attention Transformers." arXivpreprint arXiv:2306.13804. https://arxiv.org/abs/2306.13804.
[6] Ringeval, F., Sonderegger, A., Sauer, J.,& Lalanne, D. (2013). "Introducing the RECOLA Multimodal Corpus of RemoteCollaborative and Affective Interactions." In 10th IEEE InternationalConference on Automatic Face and Gesture Recognition (FG). IEEE.
[7] Alex, S. B., Mary, L., & Babu, B. P. (2020). "Attention and Feature Selection for Automatic Speech Emotion Recognition Using Utterance and Syllable-Level Prosodic Features." Circuits, Systems, and Signal Processing, 39(11), 5681-5709.
[8] Koduru, A., Valiveti, H. B., & Budati, A. K. (2020). "Feature Extraction Algorithms to Improve the SpeechEmotion Recognition Rate."International Journal of Speech Technology, 23(1), 45-55.
[9] Chen, M., He, X., Yang, J., & Zhang, H. (2018). "3-D Convolutional RecurrentNeural Networks with Attention Modelfor Speech Emotion Recognition." IEEE Signal Processing Letters, 25(10), 1440-1444.
[10] Noroozi, F., Marjanovic, M., Njegus, A., Escalera, S., & Anbarjafari, G. (2017). "Audio-Visual Emotion Recognition in Video Clips." IEEE Transactions on Affective Computing. https://doi.org/10.1109/TAFFC.2017.2777804.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

# INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY