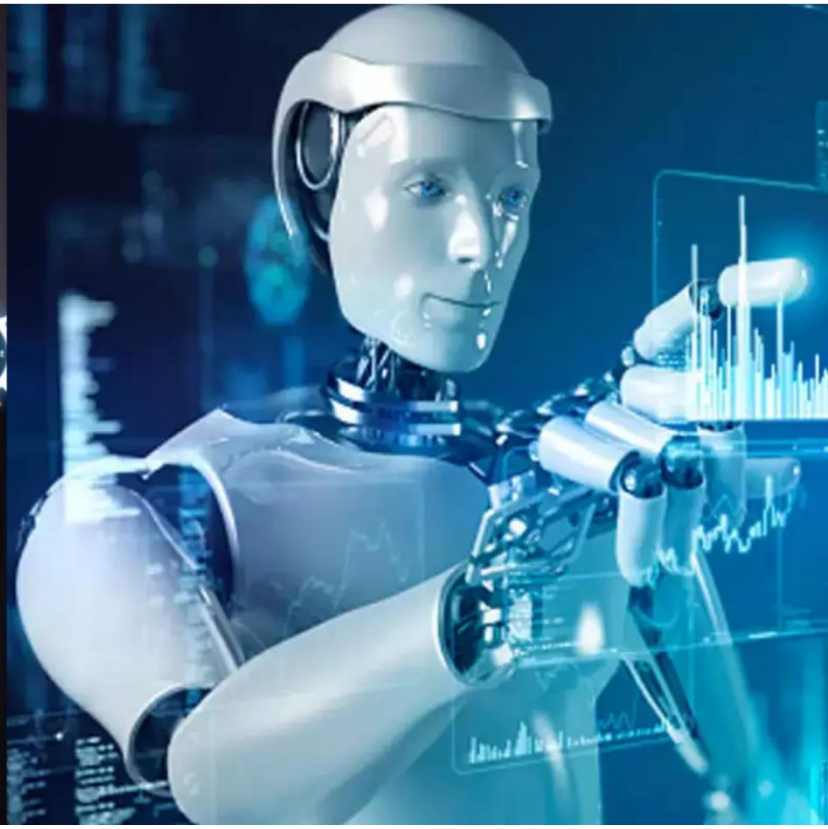
IJIRSET
International Journal of Innovative Research in Science | Engineering | Technology

ISSN(O): 2319-8753
ISSN(P): 2347-6710

International Journal of Innovative Research in Science Engineering and Technology (IJIRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)



Impact Factor: 8.699

Volume 14, Issue 3, March 2025

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 3, March 2025

|DOI: 10.15680/IJIRSET.2025.1403289|

Speech Emotion Recognition using Hybrid Methods

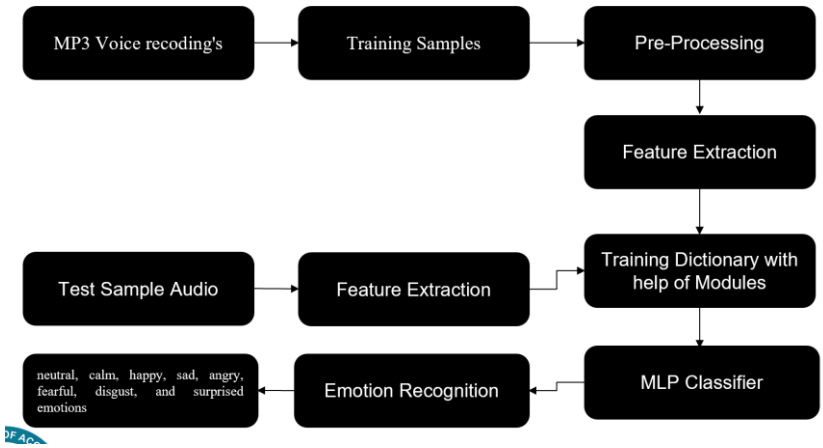
Dr. Thirupurasundari D.R, Ramireddy Bhanu Prakash Reddy, Rasamsetty Vinay Babu, Ravipati Sai Prudhvi, Rekha Venkat Manideep Reddy

Associate Professor, Department of CSE, Bharath Institute of Higher Education and Research, Selaiyur, Chennai, Tamil Nadu, India

B. Tech Students, Bharath Institute of Higher Education and Research, Selaiyur, Chennai, Tamil Nadu, India

ABSTRACT: Emotions are the most effective means for individuals to express their thoughts and actions to others. The ability to identify emotions through the voice of an individual speaker represents a significant technological milestone today. Recognizing emotions offers valuable insight into a person's mindset. The term Speech Emotion Recognition (SER) refers to the method of detecting emotions in human speech. We used the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) dataset to extract emotions from spoken language. Factors such as Mel-Frequency Cepstral Coefficients (MFCC) and Mel spectrograms are employed to identify emotions in speech. The accuracy achieved during the Multilayer Perceptron classifier (MLP) phase was 74.42%, whereas post-training with Convolutional Neural Networks and Long Short-Term Memory (CNN)(LSTM) increased the accuracy to 84.72%.

KEYWORDS: SER, Mel, MFCC, MLP, CNNLSTM.



I. INTRODUCTION

The domain of speech emotion recognition has developed into a rapidly advancing field within computer science. Emotions act as the medium through which people express their feelings and mental states. Emotional regulation holds significant importance in sensitive roles such as that of surgeons, military leaders, and various other fields. Individuals convey emotions through numerous channels, including posture, facial expressions, and vocal intonations, with speech being particularly critical for emotional communication. To enable effective interaction, systems must be adept at comprehending the emotional information portrayed through speech. Various machine learning techniques have been devised and assessed to categorize emotions, where feature extraction plays a crucial role in speech emotion recognition; the caliber of the features extracted directly influences the accuracy of classification results. Identifying emotions poses a complex challenge, and this research aims to detect the emotions present in specific audio samples, utilizing both Multilayer Perceptron and Convolutional Neural Network Long Short-Term Memory classifiers. The effectiveness of these two classifiers is evaluated, and the results indicate that the CNN LSTM model consistently predicts the emotional state of speech input more reliably than the MLP method.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 3, March 2025

|DOI: 10.15680/IJIRSET.2025.1403289|

II. LITERATURE REVIEW

Research conducted by Peipei Shen et al explored automatic speech emotion recognition using a support vector machine, achieving an accuracy of 66.02%. Seyedmahdad Mirsamadi et al developed an automatic approach for speech emotion recognition employing recurrent neural networks with nearby context, yielding an accuracy of 63.5%. Puneet Kumar et al introduced a multimodal method for recognizing four emotions from speech, utilizing gated recurrent units and bidirectional Encoder representations from transformers, resulting in an accuracy of 71%. Chi-Chun Lee et al proposed an emotion recognition technique based on a hierarchical binary decision tree aimed at identifying four emotions in speech, which showed enhancement over the SVM baseline. Mao Li et al presented a contrastive unsupervised learning method for speech emotion recognition that identifies just two emotions, specifically anger and sadness, through contrastive predictive coding.

III. METHODOLOGY

The MLP and CNN LSTM models are employed to identify emotions from speech based on the RAVDESS dataset. RAVDESS dataset contains 2800 speech files gathered from 24 actors labelled from 01 to 24. There are 12 male actors and 12 female actors among 24 actors with odd numbers representing male actors and even numbers representing female actors. Our model is trained on audio- only data since our primary objective is to identify emotions in speech. All 24 actors utter two specified statements for every one of the eight emotions repeated twice for every sentence except the "neutral" emotion, which occurs only with normal intensity. There are two intensity levels for all the emotions i.e., normal and strong. Names for the emotions are denoted by the numbers 01 to 08.

The Key characteristics of the audio data are obtained using MFCC (Mel Frequency Cepstral Coefficients) and Mel Spectrogram. MFCC is a significant feature extraction method employed when using audio datasets. Mel scale is a scale which maps the frequency of the heard tone to actual measured frequency. It scales frequency so that you can accommodate more carefully what can be heard by the human ear. Mel Spectrogram is arrived at by using a fast Fourier transform on overlapped parts of the signal, and spectrogram is a graphical mode.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 3, March 2025

|DOI: 10.15680/IJIRSET.2025.1403289|

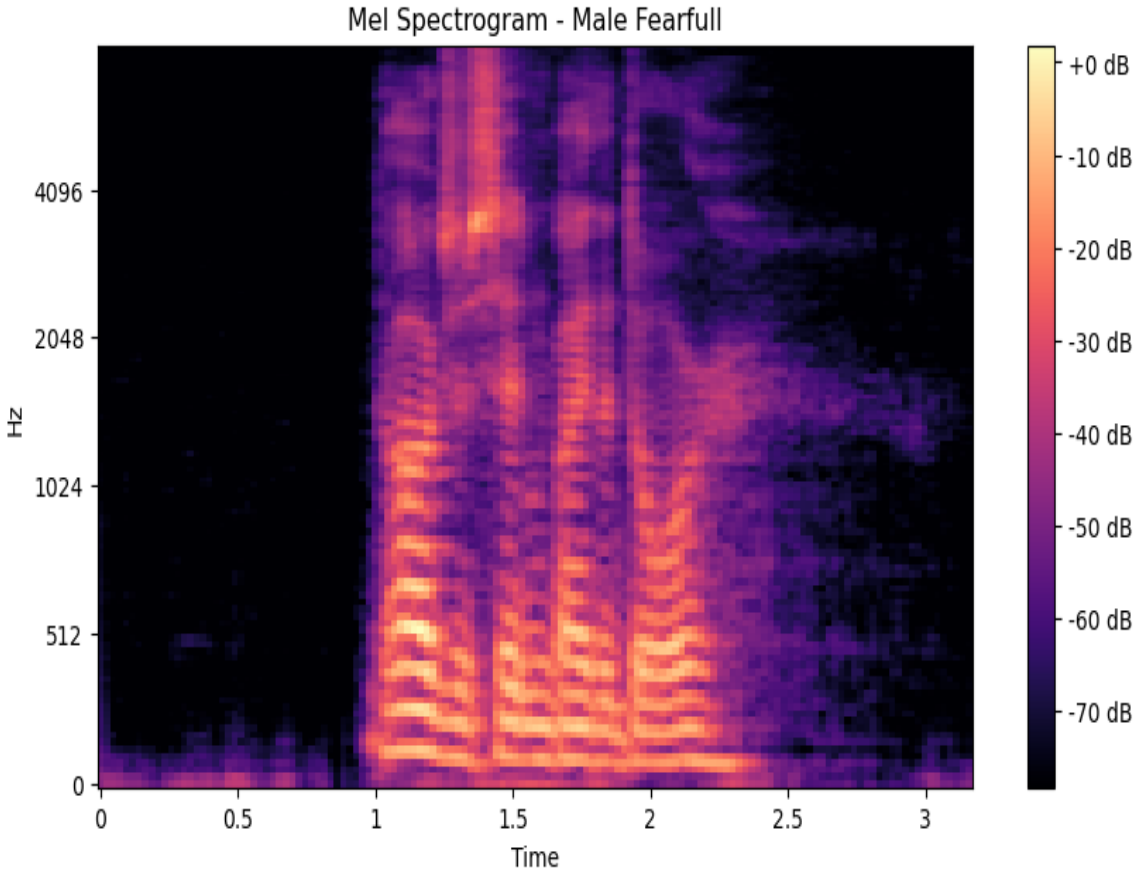


Fig 1: Mel Spectrogram for the emotion Male Fearful

A Multi-Layer Perceptron contains three types of layers. They are input layer, hidden layer and output layer. The first data is taken by the input layer and then multiplied with weights to form the hidden layer. Non-linear data is learnt by the model through activation functions in this layer. The hidden layers keep on calculating the process until the final output layer. The activation function used in our application is the rectified linear unit activation function. The hidden layer is not explicitly exposed to the input. The output layer generates a value or vector in the format required by the problem.

CNN LSTM, is an LSTM architecture that is specialized in solving sequence prediction problems. To implement a CNN LSTM, CNN layers are introduced at the start, followed by LSTM layers, and concluded with a dense layer on the output. The model can be conceptualized as having two separate parts: the CNN Model, which extracts features, and the LSTM Model, which processes the extracted features. The CNN layers extract features from the input, and LSTMs deal with sequence prediction.

IV. IMPLEMENTATION

The dataset was partitioned into training and testing subsets using the conventional 80:20 ratio. For the exploratory data analysis, MFCC and Mel Spectrogram techniques were employed. Frequency distribution of various emotion categories is depicted in the corresponding.

The training dataset is expanded through data augmentation techniques. Diverse audio samples are generated by applying various disruptions to the original training data. The aim is to improve the model's resilience to such

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 3, March 2025

|DOI: 10.15680/IJIRSET.2025.1403289|

disruptions and enhance its generalizability. More importantly, the introduced disturbances must retain the same label as the corresponding original training samples.

The features utilized in the study are cepstral coefficients, ranging from the zeroth order to the thirteenth order, along with the delta coefficients up to the fifth order, resulting in a total of 20 features. These features were extracted and compiled into a CSV file. The researchers implemented a multilayer perceptron model, incorporating two hidden layers with 512 neurons, which achieved an accuracy of 74.42 % on the test dataset. Additionally, a convolutional neural network combined with long short-term memory architecture was employed.

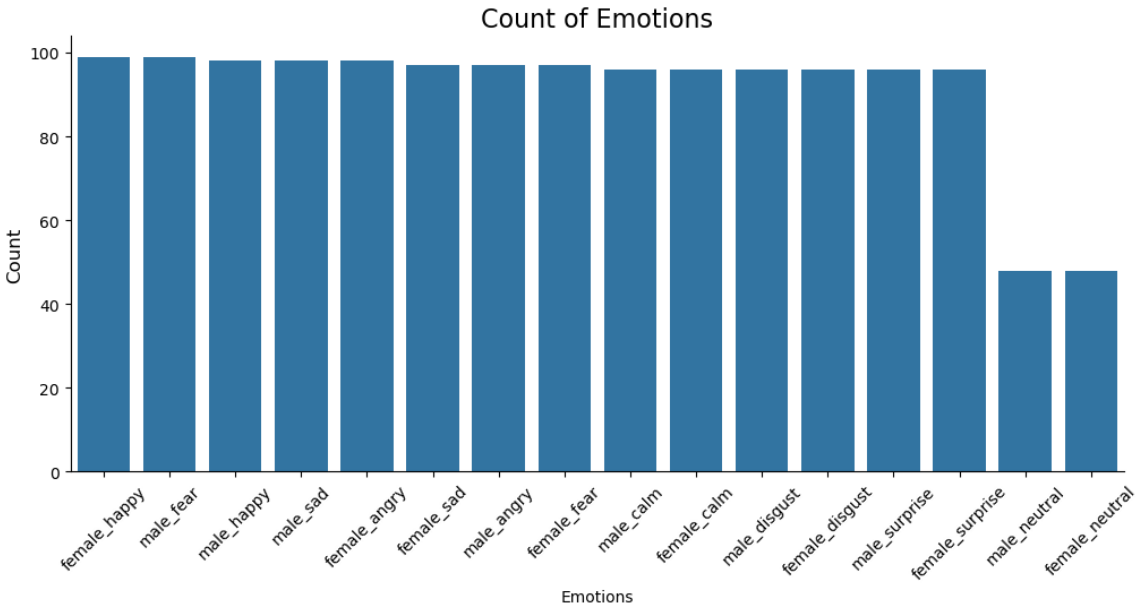


Fig 2: Total Count of Emotions

This model comprised three one-dimensional convolutional layers, each accompanied by a batch normalization layer and a max pooling layer. The activation function used for all the convolutional layers was the rectified linear unit. The network also included two LSTM layers followed by four dense layers, ending with a fully connected layer. Finally, the SoftMax function was utilized in the model.

V. RESULTS AND CONCLUSION

Speech emotions are divided into 16 classes, with 8 classes for men and 8 classes for women. On the training dataset, CNNLSTM achieved an accuracy of 84.72 percent, while MLP achieved an accuracy of 74.42 percent. Figures 3 and 4 show the CNN LSTM's confusion matrix and performance metrics.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 3, March 2025

|DOI: 10.15680/IJIRSET.2025.1403289|

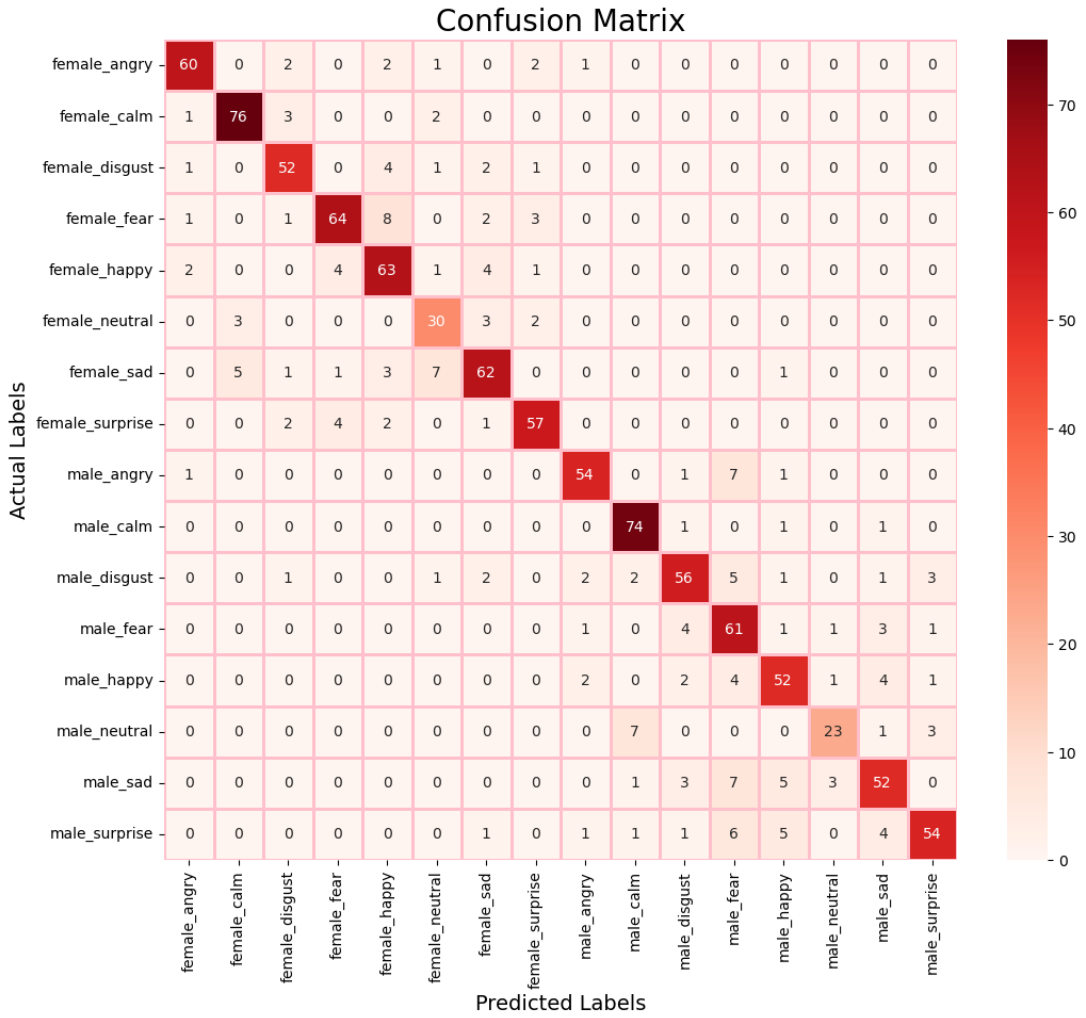


Fig 3: Confusion Matrix of CNN LSTM

For speech emotion recognition, the effectiveness of Convolutional Neural Network Long Short-Term Memory (CNN LSTM) and Multi-Layer Perceptron (MLP) models is contrasted. Through a higher accuracy rate and superiority in processing the sequential data, our experiments demonstrated that the CNNLSTM model outperforms the MLP model. According to our study, the CNN LSTM model shows the ability to solve speech emotion recognition tasks.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 3, March 2025

|DOI: 10.15680/IJIRSET.2025.1403289|

	precision	recall	f1-score	support
female_angry	0.91	0.88	0.90	68
female_calm	0.90	0.93	0.92	82
female_disgust	0.84	0.85	0.85	61
female_fear	0.88	0.81	0.84	79
female_happy	0.77	0.84	0.80	75
female_neutral	0.70	0.79	0.74	38
female_sad	0.81	0.78	0.79	80
female_surprise	0.86	0.86	0.86	66
male_angry	0.89	0.84	0.86	64
male_calm	0.87	0.96	0.91	77
male_disgust	0.82	0.76	0.79	74
male_fear	0.68	0.85	0.75	72
male_happy	0.78	0.79	0.78	66
male_neutral	0.82	0.68	0.74	34
male_sad	0.79	0.73	0.76	71
male_surprise	0.87	0.74	0.80	73
accuracy			0.82	1080
macro avg	0.82	0.82	0.82	1080
weighted avg	0.83	0.82	0.82	1080

Fig 4: Performance Metrics for CNN LSTM

VI. LIMITATIONS AND FUTURE WORK

While CNN-LSTM and MLP models can detect emotion from speech acoustics, they overlook contextual information, which is essential for understanding emotion. For example, sarcasm can be difficult to detect with these models because they require context awareness. Gestures and facial expressions can provide additional cues about emotions. In practical scenarios, the accuracy of emotion detection can be increased by integrating information from voice, facial expressions, and physiological cues.

6.1 Enhanced Data Collection and Augmentation

Gather extensive and varied datasets that encompass multiple languages, different accents, and a range of emotional expressions. Utilize sophisticated data augmentation methods, such as altering pitch and adding noise, to enhance the model's resilience.

6.2 Model Optimization and Efficiency

Investigate the use of lightweight models, like those based on attention mechanisms or Transformers, to enable real-time processing. Employ techniques like quantization and pruning to decrease model size while maintaining accuracy. Multimodal Emotion Recognition Integrate speech with other forms of data, such as facial expressions and physiological signals, to enhance emotion recognition. Apply fusion methods to combine audio and visual information for a more comprehensive contextual understanding

REFERENCES

[1] Shen, P., Changjun, Z., & Chen, X. (2011, August). Automatic speech emotion recognition using support vector machine. In Proceedings of 2011 international conference on electronic & mechanical engineering and information technology (Vol. 2, pp. 621-625). IEEE.

[2] Mirsamadi, S., Barsoum, E., & Zhang, C. (2017, March). Automatic speech emotion recognition using recurrent neural networks with local attention. In 2017 IEEE International conference on acoustics, speech and signal processing (ICASSP) (pp. 2227-2231). IEEE.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 3, March 2025

|DOI: 10.15680/IJIRSET.2025.1403289|

[3] Kumar, P., Kaushik, V., & Raman, B. (2021). Towards the Explainability of Multimodal Speech Emotion Recognition. In Interspeech (pp. 1748-1752).
[4] Lee, C. C., Mower, E., Busso, C., Lee, S., & Narayanan, S. (2011). Emotion recognition using a hierarchical binary decision tree approach. Speech Communication, 53(9-10), 1162- 1171.
[5] Li, M., Yang, B., Levy, J., Stolcke, A., Rozgic, V., Matsoukas, S., ... & Wang, C. (2021, June). Contrastive unsupervised learning for speech emotion recognition. In ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 6329-6333). IEEE.
[6] Lin, Y. P., Wang, C. H., Jung, T. P., Wu, T. L., Jeng, S.K., Duann, J. R., & Chen, J. H. (2010). EEG emotion recognition in music listening. IEEE Transactions on Biomedical Engineering, 57(7), 1798-1806.
[7] Schaaff, K., & Schultz, T. (2009, September). Towards an EEG-based emotion recognizer for humanoid robots. In RO-MAN 2009-The 18th IEEE International Symposium on Robot and Human Interactive Communication (pp. 792-796). IEEE.
[8] Pierre-Yves, O. (2003). The production and recognition of emotions in speech: features and algorithms. International Journal of Human-Computer Studies, 59(1-2), 157-183.
[9] Damodar, N., Vani, H. Y., & Anusuya, M. A. (2019). Voice emotion recognition using CNN and decision tree. Int. J. Innov. Technol. Exp. Eng, 8, 4245-4249.
[10] Zhao, J., Mao, X., & Chen, L. (2018). Learning deep features to recognise speech emotion using merged deep CNN. IET Signal Processing, 12(6), 713-721.
[11] Palo, H. K., Mohanty, M. N., & Chandra, M. (2015). Use of different features for emotion recognition using MLP network. In Computational Vision and Robotics: Proceedings of ICCVR 2014 (pp. 7-15). Springer India.
[12] Livingstone, S. R., & Russo, F. A. (2018). The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. PloS one, 13(5), e0196391.
[13] Boigne, J., Liyanage, B., & Östrem, T. (2020). Recognizing more emotions with less data using self- supervised transfer learning. arXiv preprint arXiv:2011.05585.
[14] Venkataramanan, K., & Rajamohan, H. R. (2019). Emotion recognition from speech. arXiv preprint arXiv:1912.10458.

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi cross ref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

www.ijirset.com

Scan to save the contact details