A Hybrid Approach of CNN-LSTM for Speech Emotion Recognition

Rajeev Ranjan[*1], Vivek Giri[*2], Shiv Datta Dixit[*3], Sourabh[*4]

[*1,2,3,4] Student, Department of Computer Science & Engineering, Indian Institute of Information Technology, Bhagalpur

Abstract

Speech Emotion Recognition (SER) is an emerging research field focused on categorizing speech signals into various emotional states. This area has gained prominence due to the rise of social media and the accessibility provided by the internet's low cost and high bandwidth. SER systems rely on supervised learning classifiers that are trained on labelled data, with feature extraction playing a crucial role in distilling raw data into key characteristics. To gain a comprehensive understanding of the features and techniques used for feature extraction, a detailed literature survey of previous models is included.

In this paper, we aim to advance existing research in SER by employing a hybrid approach that integrates a 1-D Convolutional Neural Network (CNN) with Long Short-Term Memory (LSTM) architecture. This combined model aims to harness the advantages of both CNN and LSTM to enhance emotion classification accuracy. We conducted experiments using a custom dataset created from the CREMA-D, TESS, RAVDESS and SAVEE databases, encompassing approximately 600 speech files per emotion. This dataset offered a rich variety of emotional expressions for thorough training and evaluation.

Keywords – CNN, LSTM, Speech Emotion Recognition

1. Introduction

[1] Machine learning (ML) is a field that encompasses the use of past experience, i.e., previous data, to enhance future performance. Its primary focus is on automated learning methods that facilitate the modification or improvement of algorithms based on prior experiences. This process occurs automatically, without the need for external assistance from humans.

[2] Speech is an essential aspect of human culture, allowing us to convey both linguistic and paralinguistic information. Although traditional automatic speech recognition systems have primarily emphasized on linguistic information, it is essential to consider paralinguistic features, including gender, personality, and emotion, to accurately detect emotions. Emotion recognition has been a topic of study for a long while, with research originally focusing on detecting emotions from facial expressions. However, recent developments have led to increasing attention on SER, as it aims to identify emotions from speech signals. Nevertheless, SER is a challenging task due to the difficulty in extracting effective emotional features from speech data.

[3] An SER system is a classification of techniques that processes and characterizes speech information to identify emotions embedded in them. Such a system requires a supervised learning classifier trained on labeled data with emotions ingrained in it. Ahead of feature extraction, the data must be preprocessed to ensure consistency in the sampling rate across all databases. To classify data accurately, the process relies on using features to distill the essential characteristics from raw data. Depending on the situation, acoustic features alone may be sufficient, or additional features

such as linguistic, facial, or speech information may be required.

[4] The performance of classifiers primarily depends on the techniques used for feature extraction and the salience of features for a particular emotion. To enhance classifiers, additional features from various modalities can be integrated, although this depends on their importance and usability. To identify emotions based on their acoustic connection in voice utterances, a wide variety of classifiers are available to pass through the characteristics. While many acoustic features and classifiers have been experimented with, accuracy still needs improvement.

In this study, we sought to advance current research in Speech Emotion Recognition (SER) by implementing a hybrid model combining a 1-D Convolutional Neural Network (CNN) with Long Short-Term Memory (LSTM) architecture. CNNs have shown outstanding performance in various fields, including image recognition, and have recently demonstrated significant potential in SER due to their capability to automatically extract pertinent features from input data. LSTMs, a type of recurrent neural network, excel at capturing long-term dependencies in sequential data—a crucial aspect for detecting emotional patterns in speech. LSTMs use gating mechanisms to manage the flow of information, allowing them to retain and discard information as necessary.

By combining CNN and LSTM in our model, we aimed to leverage the strengths of both architectures and improve the accuracy of emotion classification. We conducted experiments using a custom dataset created using labeled datasets such as CREMA-D, TESS, RAVDESS and SAVEE, which consisted of around 600 speech files per emotion. This dataset provided a diverse range of emotional expressions for training and evaluation.

2. Related Work

Recent studies on Speech Emotion Recognition (SER) have proposed various methods to enhance emotion classification performance and accuracy. These methods utilize deep learning techniques, combining different neural network architectures to effectively capture and interpret emotional signals from speech and visual data. [5] One approach employs a multimodal system using a recurrent network, specifically a Long Short-Term Memory (LSTM) network, to process raw visual and speech data, capturing contextual information for predicting natural and spontaneous emotions. [6] Another method involves a 3-D attention-based convolutional recurrent neural network (ACRNN) that extracts log-Mels with deltas and delta-deltas as input, merging 3-D convolutional neural networks (CNNs) with LSTM and incorporating an attention layer to create utterance-level affective-salient features. Similarly, [7] a model that combines attention-based bidirectional LSTM recurrent neural networks with attention-based fully convolutional networks has been proposed for automatic SER from spectrograms. [8] Another technique addresses the temporal relationship of speech waveforms by combining frame-level speech features with attention-based LSTM recurrent neural networks. Furthermore, [9] a novel architecture called ADRNN integrates dilated CNNs, residual blocks, bidirectional LSTMs, and attention mechanisms, along with a loss function that combines SoftMax with centre loss, to enhance speech emotion recognition performance. These diverse approaches illustrate the use of various neural network architectures, feature extraction techniques, fusion strategies, and deep learning methods to advance SER and capture the intricate emotional nuances in speech and visual data.

3. Data Set

3.1 Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS)

The [10] Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) is a well-established database that encompasses both song and speech recordings to capture emotional expressions.

It is designed to be multimodal, meaning it includes visual and auditory components. The database comprises recordings from 24 professional actors, ensuring gender balance and a neutral North American accent. Each actor vocalizes statements that are matched in terms of lexical content. The speech recordings cover a range of emotions, including disgust, sadness, calm, happiness, anger, surprise, and fear. Additionally, the song recordings portray emotions such as fear, sadness, anger, calm, and each emotional expression is performed at two different levels of intensity, with an extra neutral expression included. The database provides recordings in three different formats: face-and-voice, face-only, and voice-only. These formats allow researchers to examine the impact of visual cues on emotion perception. To ensure the reliability and validity of the data, each recording was rated by 247 individuals who represented non trained research participants from North America. The ratings focused on intensity, emotional validity, and genuineness and were obtained through 10 separate evaluations for each recording. Test-retest data was also collected from 72 participants to assess the reliability of the ratings over time.

3.2 Surrey Audio-Visual Expressed Emotion (SAVEE)

The Surrey Audio-Visual Expressed Emotion (SAVEE) database comprises recordings from four native English male speakers, who were postgraduate students and researchers at the University of Surrey. Their ages ranged from 27 years to 31 years. Emotions were categorized into discrete psychological categories, including surprise, anger, happiness, disgust, fear, sadness, and neutral. The database included 15 TIMIT sentences per each emotion, which consisted of 3 common, 2 emotion-specific, and 10 generic sentences that were phonetically balanced and unique for each emotion. To maintain consistency, the 3 common and 2 emotion-specific sentences were also recorded as neutral, resulting in a total of 30 neutral sentences. Thus, each speaker contributed a total of 120 utterances to the database.

3.3 Crowd Sourced Emotional Multimodal Actors Dataset (CREMA-D)

CREMA-D is an emotional multimodal actor data set of 7,442 original clips from 91 actors. These clips were from 48 male and 43 female actors between the ages of 20 and 74 coming from a variety of races and ethnicities (African America, Asian, Caucasian, Hispanic, and Unspecified). Actors spoke from a selection of 12 sentences. The sentences were presented using one of six different emotions (Anger, Disgust, Fear, Happy, Neutral, and Sad) and four different emotion levels (Low, Medium, High, and Unspecified). Participants rated the emotion and emotion levels based on the combined audiovisual presentation, the video alone, and the audio alone. Due to the large number of ratings needed, this effort was crowd-sourced and a total of 2443 participants each rated 90 unique clips, 30 audio, 30 visual, and 30 audio-visuals. 95% of the clips have more than 7 ratings.

3.4 Toronto Emotional Speech Set (TESS)

There are a set of 200 target words were spoken in the carrier phrase "Say the word _' by two actresses (aged 26 and 64 years) and recordings were made of the set portraying each of seven emotions (anger, disgust, fear, happiness, pleasant surprise, sadness, and neutral). There are 2800 data points (audio files) in total. The dataset is organised such that each of the two female actor and their emotions are contain within its own folder. And within that, all 200 target words audio file can be found. The format of the audio file is a WAV format.

4. Proposed Approach

In [11], author Surjeet Balhara et al. proposed an approach that combines a 2- D Convolutional Neural Network (CNN) with Long Short-Term Memory (LSTM) architecture. They have used a custom dataset created using labeled datasets viz. RAVDESS and SAVEE, with around 450 speech files per emotion. This approach yielded a result of 65% as seen in Figure 1.

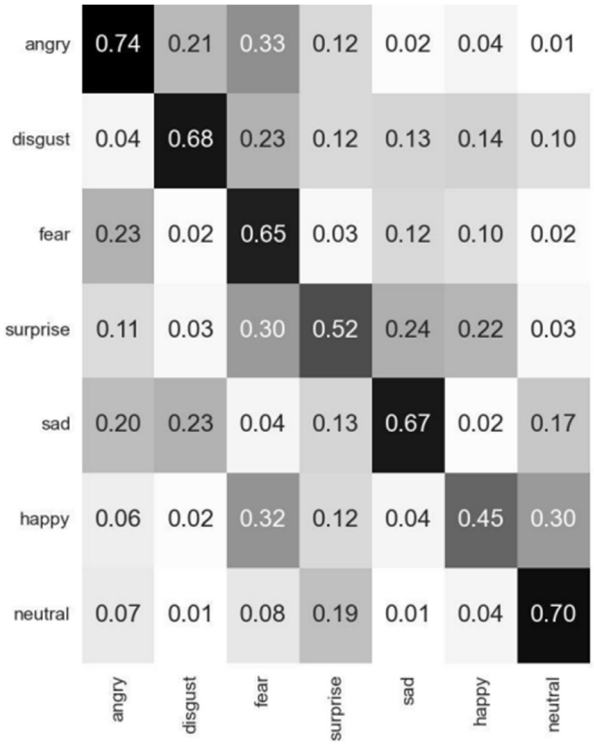

Figure 1. A confusion matrix of CNN with LSTM with an average of 65% accuracy, where each row presents the confusion of the ground truth emotion during prediction [11]

In our proposed 1D-CNN and LSTM model, the spectrograms are extracted from each audio file and enhanced before processing them using the CNN and LSTM model. In comparison to [11], our model works on a different and custom dataset created using the CREMA-D, TESS, RAVDESS and SAVEE datasets. We have explained the step-wise details of our approach towards SER.

1. Data Collection

The CREMA-D, TESS, RAVDESS and SAVEE datasets were selected for the project, as they provide a diverse range of emotional speech data. The datasets contain several folders, each representing a different emotion. The audio files in these folders were used for training and evaluation.

2. Data Preprocessing

The audio files were organized and labeled with a unique identifier representing the emotion. This labeling was crucial for determining the ground truth labels during the training and evaluation phases.

3. Data Visualization and Exploration

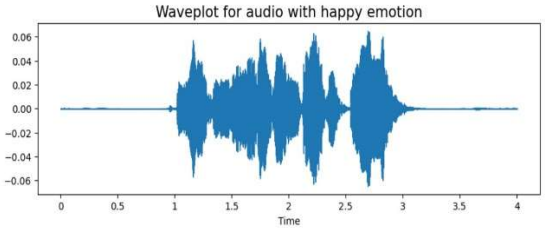

Figure 2. Wave plot for audio with happy emotion

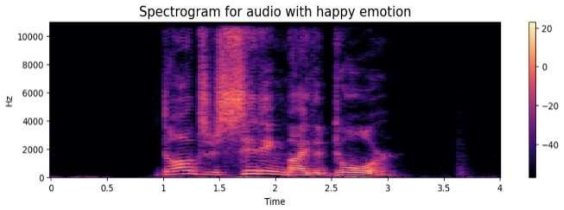

Figure 3. Spectrogram for audio with happy emotion

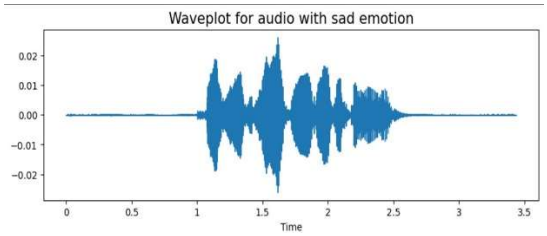

Figure 4. Wave plot for audio with sad emotion

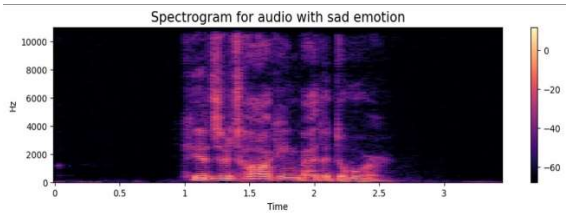

Figure 5. Spectrogram for audio with sad emotion

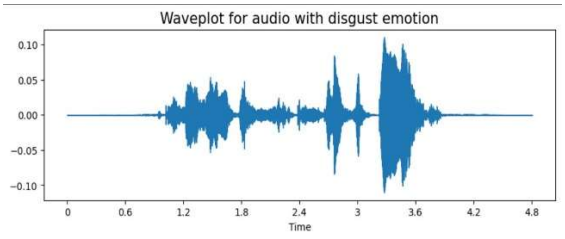

Figure 6. Wave plot for audio with sad emotion

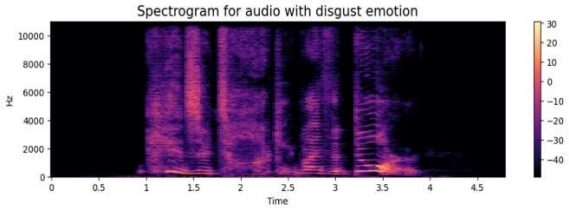

Figure 7. Spectrogram for audio with disgust emotion

4. Feature Extraction

The Python Librosa library was utilized to process and extract features from the audio files. The model focused on extracting Mel Frequency Cepstral Coefficients (MFCCs) as they are commonly used in speech and speaker recognition tasks. The extracted MFCCs provided insights into the relevant features influencing the audio data. We have used Zero-Crossing Rate which indicates the rate at which the signal changes sign. It is a measure of the noisiness or periodicity of the signal. We have also used Chroma_STFT for capturing harmonic and melodic characteristics.

5. Feature Enhancement

To optimize the extracted features, preprocessing techniques were employed. This step was designed to boost the efficiency of emotion detection. Such preprocessing methods might involve normalization, feature scaling, or noise reduction to minimize irrelevant variations and amplify significant emotional cues present in the audio data. By refining the data in this manner, the system can more accurately identify and classify emotions.

6. Model Architecture

The project utilized a combination of 1-D Convolutional Neural Network (CNN) and Long Short-Term Memory (LSTM) networks, as depicted in Figure 8. This model processes audio data represented by mel spectrograms. The CNN layers were instrumental in capturing spatial features from these spectrograms, while the LSTM layers focused on capturing temporal dependencies inherent in sequential data. CNNs have showcased outstanding performance across various fields, including image recognition, and have recently proven to be highly effective in Speech Emotion Recognition (SER). Their ability to automatically extract relevant features from input data makes them particularly useful for analyzing speech signals. LSTMs, on the other hand, are a specialized type of recurrent neural network designed to capture long-term dependencies in sequential data. This capability is crucial for SER, as emotions are often conveyed through temporal patterns in speech. LSTMs achieve this by utilizing gates to regulate the flow of information, allowing them to selectively remember and forget data as required. By integrating these two architectures, the project aims to leverage the strengths of both CNNs and LSTMs, enhancing the accuracy and robustness of emotion classification in speech signals. This combined approach has the potential to significantly improve the detection and interpretation of emotional cues in audio data.

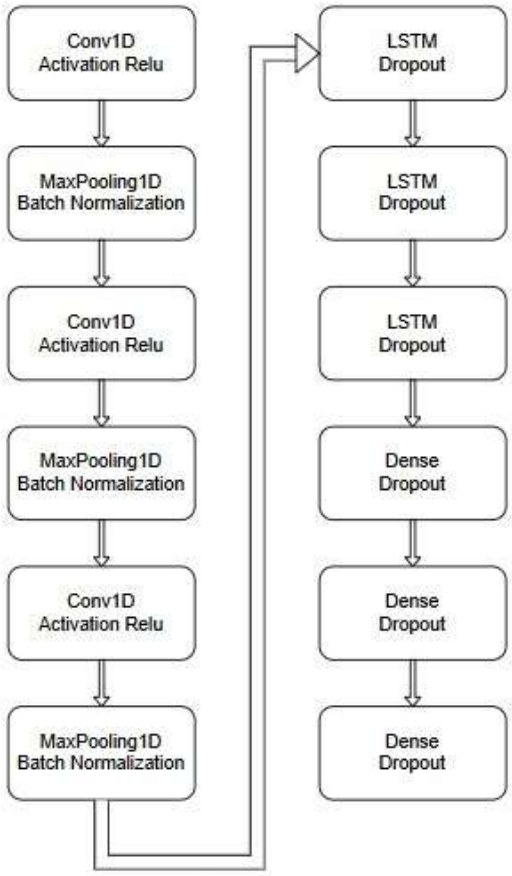

Figure 8. The CNN-LSTM baseline architecture utilized to identify voice utterances according on their emotional states.

7. Training and Evaluation

The dataset was split into training and testing sets with the model being trained on the training set using appropriate optimization algorithms and loss functions. The trained model was then evaluated on the testing set to measure its performance in recognizing speech emotions.

5. Result and Discussion

The outcomes of this study slightly surpassed those achieved by earlier approaches in the domain. Previous research has produced comparable findings. This study focused on eight distinct emotions: anger, disgust, fear, surprise, sadness, happiness, neutral, and calm. As illustrated in Figure 10, the detection accuracy improved by 16% compared to the results obtained with the 2D CNN-LSTM model [11].

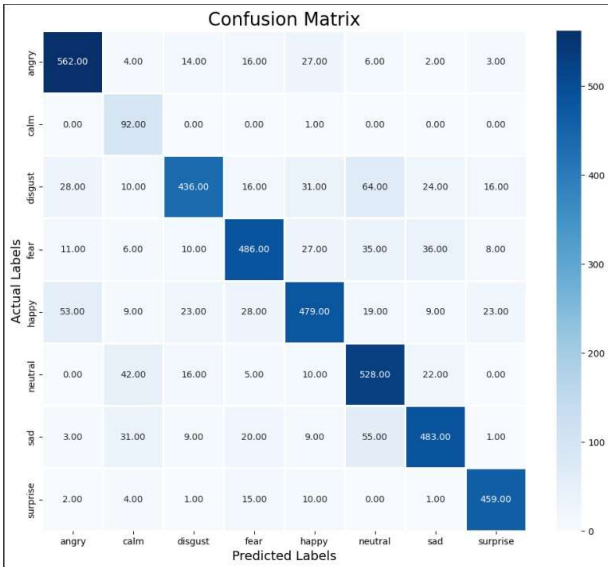

Figure 9. A confusion matrix of CNN with LSTM with an average of 81% accuracy

	precision	recall	f1-score	support
angry	0.85	0.89	0.87	634
calm	0.46	0.99	0.63	93
disgust	0.86	0.70	0.77	625
fear	0.83	0.79	0.81	619
happy	0.81	0.74	0.77	643
neutral	0.75	0.85	0.79	623
sad	0.84	0.79	0.81	611
surprise	0.90	0.93	0.92	492
accuracy			0.81	4340
macro avg	0.79	0.83	0.80	4340
weighted avg	0.82	0.81	0.81	4340

Figure 10. Classification Report

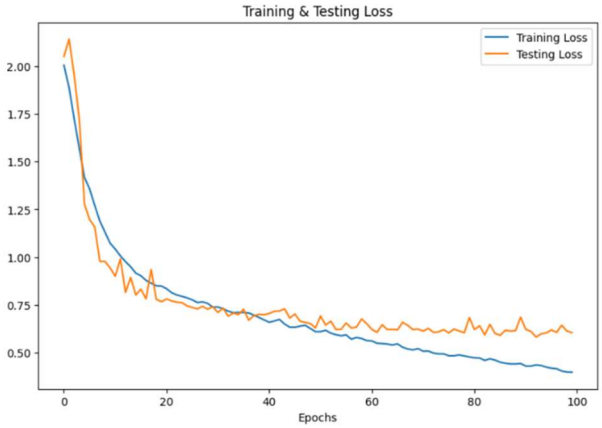

Figure 11. Training & Testing Loss

Figure 12. Training & Testing Accuracy

The accuracy in detecting calm and neutral emotions was satisfactory, suggesting that there is room for improvement with a more diverse dataset. The project achieved an overall accuracy of approximately 81%, highlighting the potential for further refinement. Future efforts should concentrate on enhancing the model's precision and minimizing errors. Addressing these issues could significantly advance the field of speech emotion recognition, increasing the reliability and accuracy of emotion detection in speech-based applications.

Although the results did not fully meet initial expectations, they provide a solid foundation for future enhancements. The project's objective was to develop a speech emotion recognition model using deep neural networks, specifically by combining Convolutional Neural Networks (CNN) with Long Short-Term Memory (LSTM) architectures. This approach was distinct from the 2-D CNN-LSTM methods [11] previously used in other studies. The proposed model was tested on a custom dataset compiled from four

databases: CREMA-D, TESS, RAVDESS, and SAVEE, each containing approximately 600 speech files per emotion.

6. Conclusion

In summary, this paper delved into Speech Emotion Recognition (SER) using Deep Neural Networks (DNN), aiming to enhance existing methodologies and tackle challenges related to language independence and the recognition of a broader range of emotions. The achieved results were slightly superior to previous methods, demonstrating advancement in the field. Although the language-independent model and the inclusion of a wider range of emotions were not fully achieved, significant progress was made towards addressing these issues.

During the implementation phase, the use of different language databases posed challenges and caused confusion for the model, underscoring the necessity for standardized databases or additional preprocessing techniques to manage linguistic variations. Another obstacle encountered was the lack of fully labelled datasets, which further complicated the model's training process.

Future research should focus on creating a custom dataset with comprehensive labelling to enhance the model's performance and accuracy. Despite some unresolved challenges, this study has made notable progress in SER using DNN. The results indicate a high potential for the model, and future efforts to develop a custom dataset and incorporate it into the training process could lead to further improvements and more precise emotion recognition.

7. References

[1] S. Balhara, N. Gupta, A. Alkhayyat, I. Bharti, R. Malik, S. Mahmood, F. Abedi. “A survey on deep reinforcement learning architectures, applications and emerging trends”, IET Communications. 2022. 10.1049/cmu2.12447.

[2] A.A Anthony, C.M. Patil, “Speech Emotion Recognition Systems: A Comprehensive Review on Different Methodologies”, Wireless Pers Commun 130, 515–525 2023. https://doi.org/10.1007/s11277-023-10296-5

[3] R. N. Behera, K. Das, “A Survey on Machine Learning: Concept, Algorithms and Applications”, International Journal of Innovative Research in Computer and Communication Engineering, Vol. 5, Issue 2, February 2017.

[4] IBM, “AI vs. Machine Learning vs. Deep Learning vs. Neural Networks: What’s the Difference?”, Available at: https://www.ibm.com/cloud/blog/ai-vs-machine-learning-vs-deep-learning-vs-neural-networks, Accessed on 24 November 2022

[5] P. Tzirakis, G. Trigeorgis, M. A. Nicolaou, B. W. Schuller and S. Zafeiriou, "End-to-End Multimodal Emotion Recognition Using Deep Neural Networks", IEEE Journal of Selected Topics in Signal Processing, vol. 11, no. 8, pp. 1301-1309, Dec. 2017, doi: 10.1109/JSTSP.2017.2764438.

[6] M. Chen, X. He, J. Yang and H. Zhang, "3-D Convolutional Recurrent Neural Networks with Attention Model for Speech Emotion Recognition", IEEE Signal Processing Letters, vol. 25, no. 10, pp. 1440-1444, Oct. 2018, doi: 10.1109/LSP.2018.2860246.

[7] J. Santoso, T. Yamada, K. Ishizuka, T. Hashimoto and S. Makino, "Speech Emotion Recognition Based on Self-Attention Weight Correction for Acoustic and Text Features", IEEE Access, vol. 10, pp. 115732-115743, 2022, doi: 10.1109/ACCESS.2022.3219094.

[8] Y. Xie, R. Liang, Z. Liang, C. Huang, C. Zou and B. Schuller, "Speech Emotion Classification Using Attention-Based LSTM", IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 27, no. 11, pp. 1675-1685, Nov. 2019, doi: 10.1109/TASLP.2019.2925934.

[9] H. Meng, T. Yan, F. Yuan and H. Wei, "Speech Emotion Recognition From 3D Log-Mel Spectrograms with Deep Learning Network", IEEE Access, vol. 7, pp. 125868-125881, 2019, doi: 10.1109/ACCESS.2019.2938007.

[10] M. Xu, F. Zhang and W. Zhang, "Head Fusion: Improving the Accuracy and Robustness of Speech Emotion Recognition on the IEMOCAP and RAVDESS Dataset", IEEE Access, vol. 9, pp. 74539-74549, 2021, doi: 10.1109/ACCESS.2021.3067460.

[11] Pranav Dutt Tripathi, Anant Krishan Sharma, Saumya Bansal, Surjeet Balhara, “Speech Emotion Recognition Using CNN and LSTM”, IJRAR June 2023, Volume 10, Issue 2, www.ijrar.org (E-ISSN 2348-1269, P-ISSN 2349-5138)