International Research Journal of Innovations in Engineering and Technology (IRJIET)
ISSN (online): 2581-3048
Volume 9, Special Issue ICCIS-2025, pp 162-167, May-2025
https://doi.org/10.47001/IRJIET/2025.ICCIS-202526
Second International Conference on Computing and Intelligent Systems (ICCIS-2025)

Speech Emotion Recognition using Convolutional Neural Networks with Attention Mechanisms

[1]A.Poongodai, [2]Y.Nandini, [3]T.Mounika, [4]A.Karishma, [5]N.Kevalya Kumar

[1]Assistant Professor, Department of CSE (AI), Madanapalle Institute of Technology & Science (Autonomous), Madanapalle, India
[2,3,4,5]Student, Department of CSE (AI), Madanapalle Institute of Technology & Science (Autonomous), Madanapalle, India
E-mails: [1]a.poongodai@gmail.com, [2]nanduroyal6@gmail.com, [3]mounimahandaiah@gmail.com, [4]alurikarishma39@gmail.com, [5]kevalyakumar@gmail.com

***Abstract* - Speech Emotion Recognition (SER) is a crucial component in enhancing human- computer interaction by enabling machines to recognize and respond to human emotions effectively. This study proposes a novel SER framework using Convolutional Neural Networks (CNNs) augmented with attention mechanisms. The CNNs are employed to capture hierarchical and spatial features from spectrogram representations of speech signals, while Attention mechanisms focus on emotionally salient regions, improving interpretability and accuracy. The proposed model is evaluated on benchmark datasets, demonstrating superior performance compared to traditional methods. This innovative combination of CNNs and attention mechanisms highlights its potential for advancing real-world SER applications such as virtual assistants, customer support systems, and mental health monitoring. By prioritizing critical emotional features, the model improves its practical utility and reliability. This work underlines the importance of deep learning techniques in developing SER technologies, paving the way for more intuitive and effective human-computer interactions. This approach highlights the potential of combining CNNs with attention for advancing SER applications in real-world scenarios.**

Keywords: Speech Emotion Recognition, Deep Learning, Convolutional Neural Networks, Attention Mechanisms.

I. INTRODUCTION

Speech Emotion Recognition (SER) is a cornerstone of affective computing, dedicated to interpreting human emotions such as joy, sorrow, anger, or fear through vocal cues. This field is significant for enabling machines to understand emotional nuances, fostering applications in human- computer interaction, psychological assessment, virtual assistants, and automated customer service. SER hinges on the premise that emotions manifest distinctly in speech through acoustic properties like pitch, intensity, rhythm, and timbre. These properties, embedded in the temporal and spectral domains of audio signals, serve as the foundation for distinguishing emotional states. The framework of SER involves a multi-stage process: extracting meaningful features from raw audio, modelling these features with advanced algorithms, and classifying them into discrete emotional categories.

The first stage, feature extraction, transforms audio waveforms into compact representations that encapsulate emotional cues. Features such as frequency-based descriptors capture the spectral envelope, reflecting timbral qualities, while temporal features track dynamic changes like speech rate or energy fluctuations. Harmonic features encode pitch relationships, and energy-based metrics highlight intensity differentiation. These features are designed to be robust to noise and speaker variability, ensuring generalizability across diverse datasets. Data augmentation plays a role by introducing variations such as pitch shifts or tempo changes to simulate real-world diversity, enhancing model robustness by exposing it to altered but plausible emotional expressions. Once extracted, features are processed by machine learning models, with deep neural networks being the gold standard due to their ability to learn complex, non-linear patterns. Convolutional Neural Networks (CNNs) excel at detecting local patterns within feature representations, hierarchically constructing abstract features that capture emotional signatures.

For instance, CNNs can identify frequency modulations or energy bursts indicative of specific emotions. To account for the sequential nature of speech, recurrent neural networks, particularly Long Short-Term Memory (LSTM) units, model temporal dependencies, preserving context across extended audio segments. Bidirectional LSTMs enhance this by considering both past and future contexts, offering a richer understanding of emotional flow. Attention mechanisms further refine the process by assigning weights to different time steps, emphasizing segments of speech that carry stronger emotional signals, such as stressed syllables or intonation shifts. This selective focus improves classification accuracy by

International Research Journal of Innovations in Engineering and Technology (IRJIET)
ISSN (online): 2581-3048
Volume 9, Special Issue ICCIS-2025, pp 162-167, May-2025
https://doi.org/10.47001/IRJIET/2025.ICCIS-202526
Second International Conference on Computing and Intelligent Systems (ICCIS-2025)

filtering out less relevant information. Skip connections, another construct, integrate features from earlier layers with later ones, ensuring that low-level acoustic details are not lost during deep processing. Together, these components form a hybrid architecture that balances spatial and temporal modelling, categorical cross entropy, to minimize the discrepancy between predicted and true emotion labels. Gradient-based optimizers, such as Adam, adjust model parameters iteratively, guided by backpropagation. Regularization technique like dropout, batch normalization, and early stopping are essential to prevent overfitting, especially given the variability in emotional expression across speakers, cultures, and contexts. Hyper-parameter tuning, including learning rates and layer sizes, further refines performance. The objective is to achieve high accuracy and generalizability, enabling the model to handle unseen data effectively.

Evaluation in SER focuses on metrics like accuracy and loss, with validation on separate test sets ensuring unbiased performance estimates. Visualizations of training dynamics, such as accuracy and loss curves, provide insights into model convergence and overfitting risks. The ultimate goal is to develop a system that not only classifies emotions accurately but also generalizes to real-world scenarios, where emotional expressions may be subtle, overlapping, or context-dependent. By synthesizing robust feature extraction, advanced neural architectures, and principled training strategies, SER aims to bridge the gap between human emotional communication and computational intelligence, paving the way for empathetic, responsive technologies that the enhance human-machine interactions across diverse domains.

II. RELATED WORK

Speech Emotion Recognition (SER) is a process that involves identifying and interpreting human emotions from spoken language. It combines techniques from signal processing, machine learning, and psychology to analyse vocal attributes such as pitch, tone, tempo, and energy to determine the speaker's emotional state. The goal of SER systems is to automatically recognize emotions like happiness, anger, sadness, fear, surprise, or neutrality from audio signals, enabling more natural and empathetic human-computer interactions. SER has applications in various fields, including virtual assistants, customer service, healthcare, and affective computing.

Khalil et. al. 2019 [1] The paper "Speech Emotion Recognition Using Deep Learning Techniques: A Review" provides a comprehensive overview of how deep learning has transformed the field of speech emotion recognition (SER). Aouani & Ben Ayed 2020 [2] The study focuses on improving Speech Emotion Recognition (SER) through a hybrid approach combining handcrafted feature extraction with deep feature selection. Kaur, Jasmeet & Anil Kumar 2021 [3] The study explores the application of various machine learning algorithms—CNN, k-Nearest Neighbors (k-NN), Multi-Layer Perceptron (MLP), and RandomForest—for Speech Emotion Recognition using the Berlin Database of Emotional Speech.

Anastasia Pentari 2024 [4] The study proposed a graph-based feature extraction method for Speech Emotion Recognition, utilizing structural and statistical adjacency matrices derived from visibility graph transformations of speech signals. Apoorva Sharma, Himanshu Nawani, Shalini Verma 2023, The study presented a deep learning- based Speech Emotion Recognition (SER) model using MFCC features combined with CNN, LSTM, and a hybrid CNN+LSTM architecture.[5] Pavithra et. al. 2023 [6] This study proposed a deep learning-based Speech Emotion Recognition (SER) model using a Convolutional Neural Network (CNN) architecture, tested on the RAVDESS dataset. The model achieved a strong accuracy of 89.8%.

Congshan Sun, Haifeng Li, Lin Ma 2023 [7] This study introduced an enhanced Speech Emotion Recognition (SER) model. This approach addresses limitations of traditional EMD techniques, such as residual noise and fixed parameter settings, though it still faces challenges like high computational cost. D. Lakshmi et al. 2023 [8] This study explored various deep learning techniques for Speech Emotion Recognition (SER), including DNN, DBN, CNN, RNN, LSTM, and Autoencoders, using datasets such as IEMOCAP, EmoDB, and SAVEE

Yunhao Zhao et al. 2023 [9] This study proposed a Speech Emotion Recognition (SER) method combining spectro-temporal modulation (STM) and entropy feature extraction with CNN and classification techniques such as Gamma Classifier (GC) and Error-Correcting Output Codes (ECOC) highlighting the complexity of multi-class emotion recognition. Samarth Adkitte et al. 2023 [10] This study focused on enhancing emotion detection through feature optimization, demonstrating improved performance in identifying emotional states from speech signals.

Tae-Wan Kim and Keun-Chang Kwak 2024 [11] This study proposed a transfer learning-based Speech Emotion Recognition (SER) approach using explainable techniques, Gaussian data selection, and preprocessing with Short-Time Fourier Transform (STFT) and achieved an 87% classification accuracy. Francesco Ardan Dal Ri, Fabio Cifariello Ciardi, and Nicola Conci 2023 [13] This study utilized a CNN-based architecture with Convolutional Attention Blocks for Speech Emotion Recognition, incorporating data augmentation

International Research Journal of Innovations in Engineering and Technology (IRJIET)
ISSN (online): 2581-3048
Volume 9, Special Issue ICCIS-2025, pp 162-167, May-2025
https://doi.org/10.47001/IRJIET/2025.ICCIS-202526
Second International Conference on Computing and Intelligent Systems (ICCIS-2025)

techniques during training. The model was tested on the RAVDESS, TESS, CREMA-D, and IEMOCAP datasets, achieving accuracies of 83%, 100%, 68%, and 63%, respectively, and demonstrated generalization through cross-validation. However, the study highlighted challenges in generalizing across different datasets, as the high variability in datasets reduced the model's adaptability to real-life scenarios.

In Speech Emotion Recognition, Convolutional Neural Networks (CNNs) are powerful tools for analysing audio features, especially those represented visually like spectrograms. However, not all parts of a speech signal contribute equally to emotion recognition. To improve performance, attention mechanisms are combined with CNNs to guide the model's focus toward the most emotionally relevant segments of the input. This approach allows the system to weigh important features more heavily while reducing the influence of less informative ones. As a result, CNNs enhanced with attention can better capture the emotional nuances in speech, leading to higher recognition accuracy and more robust emotion classification across different speakers and speaking styles.

III. METHODOLOGY

3.1 Overview

Speech and Emotion refers to a method for creating realistic, expressive facial animations that align seamlessly with spoken or sung audio, enriched by emotional cues. This technique aims to simulate natural lip movements, facial expressions, and head gestures that reflect the emotions conveyed in speech or song. The primary challenge lies in accurately synchronizing audio-driven facial expressions, head movements, and emotional states in a way that feels lifelike. By leveraging advanced machine learning models, such as convolutional neural networks (CNNs) and attention mechanisms, these methods decompose complex audio inputs into meaningful components. These components, such as vocal flow and emotional tone, are used to drive realistic facial animations.

For example, a joyful tone might result in wider smiles and dynamic head movements, while a melancholic tone could induce subtle gestures and softer expressions.

One innovative approach involves decoupling and fusing different elements of the input audio such as human voice and background music. The use of an attention mechanism enables the model to focus on critical audio features that contribute most to emotional expression. This helps in generating nuanced facial movements that align with both the spoken words and the mood conveyed.

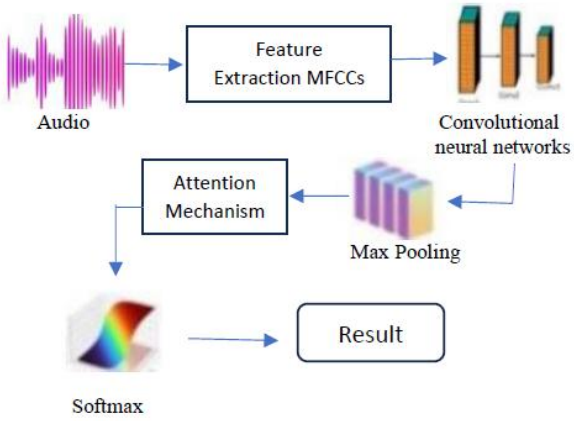

Figure 1: Block Diagram

The expressiveness is further enhanced by addressing micro-expressions and natural behaviours like blinking and head tilts. Blinks, for instance, can be modelled as short-term (frequent blinking) or long- term (prolonged eye closure), reflecting states like surprise or relaxation. Similarly, head movements are designed to match the tempo and direction of the speech or song, adding depth to the realism. This dataset includes diverse singing and speaking samples paired with synchronized facial recordings, enabling robust model training. The experiments conducted with these methods consistently show superior results in generating realistic this synthesis technique has potential applications in virtual avatars, entertainment, and human-computer interaction.

3.2 Dataset

The RAVDESS (Ryerson Audio-Visual Database of Emotional Speech and Song) is a well-known, multi- modal dataset designed for research in emotional speech and facial expression recognition. It contains 24 professional actors (12 male, 12 female) vocalizing two lexically-matched statements in a range of emotions. These emotions include neutral, calm, happy, sad, angry, fearful, surprised, and disgusted. Each emotion is displayed in two different intensity levels, with the exception of the neutral state, which is represented by a single level.

The RAVDESS dataset comprises 1,440 recordings, with each utterance carefully labelled and validated through crowd-sourced ratings to ensure emotional accuracy Researchers often choose RAVDESS for its balance between emotional variety and technical quality, making it a benchmark in affective computing research. The dataset provides a rich resource for training and evaluating models in emotion recognition tasks across different modalities, including speech, visual, and combined audio-visual analysis. The filename consists of a 7-part numerical identifier (e.g., 03-01-06-01-02-

International Research Journal of Innovations in Engineering and Technology (IRJIET)
ISSN (online): 2581-3048
Volume 9, Special Issue ICCIS-2025, pp 162-167, May-2025
https://doi.org/10.47001/IRJIET/2025.ICCIS-202526
Second International Conference on Computing and Intelligent Systems (ICCIS-2025)

01-12.wav). These identifiers define the stimulus characteristics:

1. Modality (01 = full-AV, 02 = video-only, 03 = audio-only).
2. Vocal channel (01 = speech, 02 = song).
3. Emotion (01 = neutral, 02 = calm, 03 = happy, 04 = sad, 05 = angry, 06 = fearful, 07 = disgust, 08 = surprised).
4. Emotional intensity (01 = normal, 02 = strong).
NOTE: There is no strong intensity for the 'neutral' emotion.
5. Statement (01 = "Kids are talking by the door", 02 = "Dogs are sitting by the door").
6. Repetition (01 = 1st repetition, 02 = 2nd repetition).
7. Actor (01 to 24. Odd numbered actors are male, even numbered actors are female).

Table 1: Samples per class

CLASSES	RAVDESS
Angry	192
Happiness	192
Sadness	192
Neutral	96
Disgust	192
Fear	192
Surprised	192
Calm	192

IV. EXPERIMENTAL SETUP

4.1 Validation pipeline:

The validation pipeline for a Speech Emotion Recognition system using CNN with attention mechanisms involves several structured steps to ensure the model's performance is reliable and generalizable. After training the model on labelled emotional speech data, the validation process begins by feeding unseen validation samples through the same preprocessing steps, such as noise reduction, feature extraction (e.g., spectrogram or MFCC generation), and normalization. These features are then passed through the trained CNN layers, which extract hierarchical representations of the emotional content. The attention mechanism then dynamically highlights the most relevant features, allowing the model to concrete on emotionally significant parts of the input. The final output is compared with the true emotion labels by employing evaluation metrics like accuracy, precision, recall, and the F1- score. This pipeline helps identify overfitting, tune hyperparameters, and refine model architecture, ensuring that the system performs well not only on training data but also on new, real-world audio inputs.

This method splits the dataset into multiple subsets, training the model on different combinations while validating on the remaining parts, making sure the model performs reliably across different segments of the data. During validation, loss curves and confusion matrices are analyzed to observe how well the model differentiates between emotions and to detect any misclassifications. Adjustments such as modifying CNN filter sizes, tuning attention weights, or applying regularization techniques may be made based on validation feedback. This iterative process helps in optimizing the model's structure and achieving better generalization, ultimately resulting in more precise and reliable emotion detection from speech signals.

V. RESULT

Model achieved an accuracy of 88% on the test dataset, demonstrating its capability for effective speech emotion recognition. The architecture utilized a combination of Convolutional Neural Networks (CNNs) and Bidirectional Long Short-Term Memory (BiLSTM) layers, enhanced with an attention mechanism, to extract meaningful features and model temporal dependencies. This hybrid design ensured that both local and sequential pattern in the data were effectively captured.

Data augmentation techniques, including pitch shifting and time-stretching, played a crucial role in improving the model's generalization by introducing variations into the training set. Features such as MFCCs, chroma, mel spectrograms, spectral contrast, and tonnetz offered a rich representation of audio signals, incorporating spectral, harmonic, and temporal properties. The augmentation process ensured the robustness of the model on unseen data. The CNN layers extracted local features from the input while pooling layers reduced dimensionality, preserving essential information.

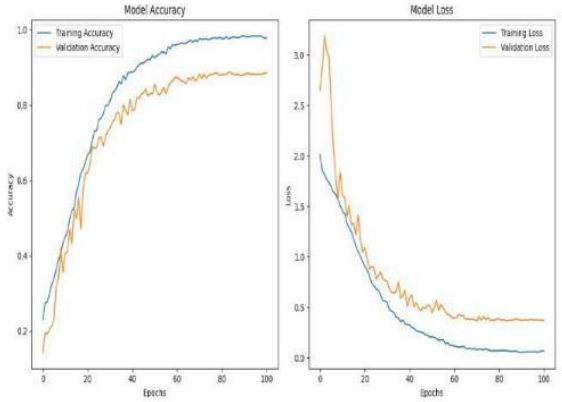

Figure 2: Model accuracy and model loss

Regularization techniques like dropout and batch normalization were employed to mitigate overfitting and stabilize the training process. Skip connections were used to retain critical information, which was further processed by the BiLSTM layers. BiLSTM layers captured both forward and

International Research Journal of Innovations in Engineering and Technology (IRJIET)
ISSN (online): 2581-3048
Volume 9, Special Issue ICCIS-2025, pp 162-167, May-2025
https://doi.org/10.47001/IRJIET/2025.ICCIS-202526
Second International Conference on Computing and Intelligent Systems (ICCIS-2025)

backward dependencies in the audio sequences, critical for understanding the temporal structure of speech. The attention mechanism added to these layers allowed the model to focus on key features, improving interpretability and relevance. The attention enhanced output was normalized with layer normalization, ensuring consistent gradient flow and stability during training.

GELU activation functions enabled smooth and effective training. The model was optimized using the Adam optimizer with a learning rate of 0.0002, and training callbacks like early stopping and learning rate reduction ensured efficient convergence. The consistent alignment between training and validation loss and accuracy indicated that the model generalized well without overfitting. The evaluation results highlighted the effectiveness of advanced feature extraction techniques, such as combining MFCCs with delta and delta-delta components and leveraging tonal features like chroma and spectral contrast. The final dense layers translated learned features into accurate emotional predictions, supported by dropout regularization for improved robustness.

VI. CONCLUSION

In this research, system presented utilizes an advanced combination of feature extraction, data augmentation, and deep learning techniques to classify speech emotions. The RAVDESS dataset forms the backbone of this study, comprising diverse speech samples categorized into eight emotions such as happiness, sadness, and anger. Feature extraction is central to the approach, with the librosa library employed to derive key audio features. These include MFCCs (mel-frequency cepstral coefficients) for capturing spectral properties, chroma features for harmonic content, mel spectrograms for spectral envelopes, spectral contrast to distinguish frequency bands, and tonnetz for tonal characteristics. Data augmentation enhances the robustness and variability of the model by applying methods like pitch modulation and time distortion applied to audio samples. The integration of CNN, BiLSTM, and attention layers achieves significant accuracy improvements in classifying emotions, validating the effectiveness of the approach.

REFERENCES

[1] Khalil et al., Edward jones. Speech Emotion Recognition using Deep Learning Techniques https://ieeeaccess.ieee.org/
[2] Aouani & Ben Ayed, Yassine Ben Ayed (2020). Speech Emotion Recognition with Deep Learning https://www.sciencedirect.com/search?qs=speech%20emotion%20recognition
[3] Kaur, Jasmeet & Anil Kumar, Shwethashri k (2021). Speech Emotion Recognition using Machine Learning https://www.irjet.net/archives/V7/i9/IRJETV7I9154
[4] Anastasia Pentari, George Kafentzis, Manolis Tsiknakis (2024). Speech Emotion Recognition via graph based representation. https://www.nature.com/articles/s4159024-52989-2.
[5] Apoorva Sharma, Himanshu Nawani, Shalini Verma (2023) Speech Emotion Recognition using Deep Learning.
[6] Pavithra et al., Sukhanya Ledella, Sirisha Devi (2023). Deep Learning based Speech Emotion Recognition: An Investigation into a sustainably Emotion–speech-Relationship. http://doi.org/10.1051/e3sconf/2023430010.
[7] Congshan Sun, Haifeng Li, Lin Ma (2023). Speech Emotion Recognition based on improved masking EMD and convolutional recurrent neural network. https:/doi.10.3389/fpsyg.2022.1075624.
[8] D. Lakshmi et al., Samuel kakuba et al. (2023). Speech Emotion Recognition using Librosa using hybrid models.
[9] Yunhao Zhao et al. (2023). Speech Emotion Recognition using convolutional Neural Networks (CNN) and gamma classifier-based error correcting output codes (ECOC). http://www.nature.com/scientificreports
[10] Samarth Adkitte et al., Vina Lomte, Mansi Fale, Vaibhavi k Kudale (2023). Speech Emotion Recognitionusing Deep Learning. https://ijcrt.org/papers/IJCRT2105446.pdf
[11] Tae-Wan Kim, keun-Chang Kwak (2024). Speech Emotion Recognition using Deep Learning Transfer Models and Explainable Techniques.
[12] Francesco Ardan Dal Ri, Fabio Cifariello Ciardi, Nicola Conci (2023). Speech Emotion Recognition and Deep Learning: An Extensive Validation Using Convolutional Neural Networks.

International Research Journal of Innovations in Engineering and Technology (IRJIET)
ISSN (online): 2581-3048
Volume 9, Special Issue ICCIS-2025, pp 162-167, May-2025
https://doi.org/10.47001/IRJIET/2025.ICCIS-202526
Second International Conference on Computing and Intelligent Systems (ICCIS-2025)

Citation of this Article:

A.Poongodai, Y.Nandini, T.Mounika, A.Karishma, & N.Kevalya Kumar. (2025). Speech Emotion Recognition using Convolutional Neural Networks with Attention Mechanisms. In proceeding of Second International Conference on Computing and Intelligent Systems (ICCIS-2025), published in *IRJIET*, Volume 9, Special Issue ICCIS-2025, pp 162-167. Article DOI https://doi.org/10.47001/IRJIET/2025.ICCIS-202526

****

