OPEN ACCESS

Kashf Journal of Multidisciplinary Research
Vol: 02 - Issue 06 (2025)
P-ISSN: 3007-1992 E-ISSN: 3007-200X
https://kjmr.com.pk

EFFECTIVE SPEECH EMOTION RECOGNITION USING R-CNN & BLSTM

Muhammad Hassan Askari
Department of Computer Science, NFC Institute of Engineering and Technology, Multan, Pakistan.

Adeel Shahzad
Department of Computer Science, Virtual University of Pakistan.

Ahmed Faraz
Department of Computer Science, University of Lahore, Pakistan.

Muhammad Fuzail*
Department of Computer Science, NFC Institute of Engineering and Technology, Multan, Pakistan.

Naeem Aslam
Department of Computer Science, NFC Institute of Engineering and Technology, Multan, Pakistan.

Mohsin Ali Tariq
Department of Computer Science, NFC Institute of Engineering and Technology, Multan, Pakistan.

****Corresponding author: Muhammad Fuzail (*** mfuzail@nfciet.edu.pk***)***

Article Info

Abstract
Speech Emotion Recognition (SER) is gaining significant attention in the field of human-computer interaction (HCI) over past decade. Specially in the fields like health, security, communication, and entertainment. But due to the lack of research on how to boost the speech processing efficiency, the current emotion recognition systems need improvement and more accuracy. To enhance the accuracy, we proposed an Effective Speech Emotion Recognition System (ESERS) which is a hybrid approach that uses Autoencoders (AEs) for denoising and robust feature extraction with a Self-Attentional Convolutional Neural Network–Bidirectional Long Short-Term Memory (CNN-BLSTM) architecture for effective temporal and contextual modeling. Using CREMA Dataset, we achieved Weighted Accuracy (WA) improved from 73.9% to 81.6% and Unweighted Accuracy (UA) increased from 68.5% to 82.8%. which shows absolute improvement of 7.7% and 14.3%, and relative improvements of 10.4% and 20.9% respectively. Hence, to enhance system efficiency, the hybrid approach outperforms traditional approaches currently in use.

Keywords:
Speech Emotion Recognition , R-CNN , BLSTM , Deep Learning , Emotion Detection , Audio Processing , Neural Networks.

Introduction

As human-computer interaction (HCI) becomes more important in everyday life, it is now more important than ever that machines understand human emotions and respond accordingly. Mostly in the recent past, people communicate with machines using a command line interface or textual input. Through these communication techniques, the emotional part of the conversation is ignored, which is very important in driving human behavior. Speech can tell you multiple things about the speaker, like their mental state, intentions, and level of engagement. These are very important for natural and intuitive communication. But it's still very hard to figure out how a person is feeling from the way of they talk because of different speakers, their accents, languages, background noise, and also the fact that people often show their feelings in a way that isn't very obvious.

As machines don't have emotions, systems like virtual assistants, educational platforms and customer service bots often respond in a very dull and unsatisfying way. As society moves toward AI systems that are smarter and more caring, making machines better at understanding emotions through Speech Emotion Recognition (SER) is emerging as a significant research topic. Figure1 shows basic SER System.

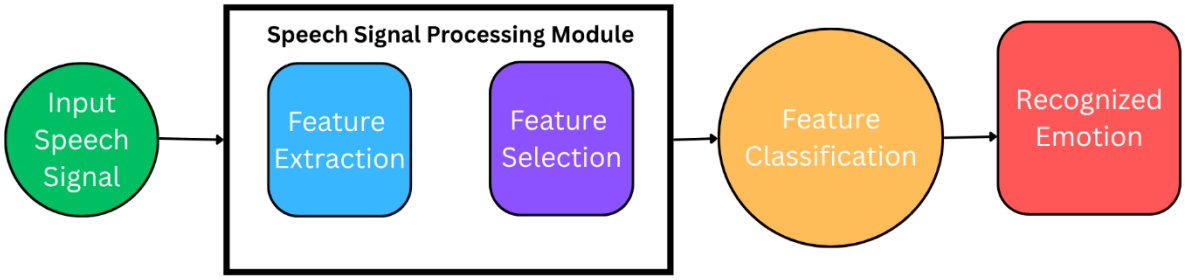

Figure 1: Basic Emotion Recognition System

Recent SER research has significantly embraced CNN–BLSTM architectures, demonstrating their strength in capturing spatial as well as temporal features from the speech. There is a study in 2021 that introduced Light-SERNet, which is a lightweight fully convolutional network that processes spectrograms and achieves high accuracy using the IEMOCAP dataset and the EMO-DB datasets by using only convolutional layers [1]. Similarly, the DCNN-BLSTM with attention model proposed by Xu et al. in 2021, which uses pretrained DCNNs (on ImageNet) for the extraction of segment-level log-Mel features, which is followed by BLSTM and attention layers, reports a huge success and gets high accuracies [2]. These latest studies are highlighting how hybrid CNN–BLSTM frameworks, especially when used together with attention features, transfer learning techniques, or pretrained representations, are currently a dominant and effective class of solutions.

Another prominent trend in the recent literature is the use of data augmentation and multi-channel or parallel feature architectures. For example, a 2023 IEEE Transactions study proposed a multichannel CNN–BLSTM model that fuses magnitude and phase spectral features (MFCC + MODGD), which is enhanced by Deep Canonical Correlation Analysis (DCCA) for feature alignment more accurately, achieving improved speaker-independent SER performance. Moreover, there is a 2024 survey by Artificial Intelligence Review which reports that if we combine MFCCs, ZCR, spectrograms, chroma and augmentation techniques like spectrogram shifting and noise addition, significantly boost CNN + BLSTM models—some reaching very high accuracy on databases like TESS, EmoDB and RAVDESS [3]. Parallel architectures, which include CNN–BLSTM–Attention networks with multi-fold augmentation, have also been proposed to address variable noise and enhance generalization on datasets like RAVDESS. These

innovations demonstrate that combining diverse data-augmentation strategies with parallel CNN and recurrent modules improves emotion recognition systems' performance as a result.

The representation of audio signals as spectrogram images in speech emotion recognition has enabled the use of advanced computer vision techniques, especially CNN-based architectures. However, traditional CNNs often face difficulties in identifying small emotional signals within the spectrograms. Recent research has focused on Region-based CNNs (R-CNN), which can identify and focus on important spectro-temporal regions, improving the emotion-specific feature extraction. When these types of models are used with networks like BLSTM, they effectively catch the temporal dynamics and spectral richness of the voice signals. For instance, recent research on SER Based on Parallel CNN-Attention Networks with Multi-Fold Data Augmentation in 2022 employs a parallel CNN–BLSTM–Attention architecture over RAVDESS and reports state-of-the-art performance [4][5]. Complementary techniques, such as multi-scale CNN with attention and co-attention fusion of MFCC, spectrogram, and wav2vec2 embeddings, demonstrate the effectiveness of mixing spatial and sequential modeling to enhance the recognition process [5]. As a result, the hybrid R-CNN + BLSTM framework shows significant results for real-world, robust and speaker-independent SER systems.

Dataset and Methodology

This paper presents a model of Recurrent Convolutional Neural Network (RCNN)-based Effective Speech Emotion Recognition (ESER) system, that uses multiple unsupervised learning techniques to improve the accuracy of identifying emotions which is speaker-independent. The techniques involve three Autoencoders (AE), Denoising AE, Adversarial AE, Variational AE, and Adversarial Variational Bayes (AVB). Experiments reveal that feature learning using unsupervised method helps improve ASER accuracy when we trained them using CEMA-D dataset, which shows improvements in UAR and also in the F1-score. The models, which include Auto encoders and AVB, show more effective performance in emotion detection compared to Denoise AE. which proves that the un-supervised learning method using these models is a valuable approach for ASER.

We used 5 different types of data sets in our research to compare the effectiveness and finding the best one for our model, firstly we used Crowd-sourced Emotional Multimodal Actors Dataset (CREMA-D), the second dataset is Toronto Emotional Speech Set (TESS), third is The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS), fourth is Danish Emotional Speech Database (DES) and the last one is Berlin Database of Emotional Speech (EMO-DB). All of their statistics are shown in Table 1 and the sample distribution throughout the databases is shown in Figure 2.

Dataset	Number of Emotions	Number of Samples	Number of Speakers	Average Length	Anger	Happiness	Sadness	Neutral	Surprise	Fear	Disgust	Boredom	Calm
Berlin Database of Emotional Speech (EMO-DB)	7	700	10	2.8 s	•	•	•	•	•	•	•	•	
Danish Emotional Speech Database (DES)	5	210	4	2.7 s	•	•	•	•					
The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS)	8	2496	24	3.7 s	•	•	•	•	•	•	•		•
Toronto Emotional Speech Set (TESS)	7	2800	2	2.1 s	•	•	•	•	•	•			
Crowd-sourced Emotional Multimodal Actors Dataset (CREMA-D)	6	7442	91	2.5 s	•	•	•	•	•	•	•		•

Table 1: Statistics of the databases and their types of emotion.

The presence of a black dot (•) in the cell shows that this specific emotion is available in that dataset. The area in the table, which is light gray in color, displays that these emotions are common in all datasets.

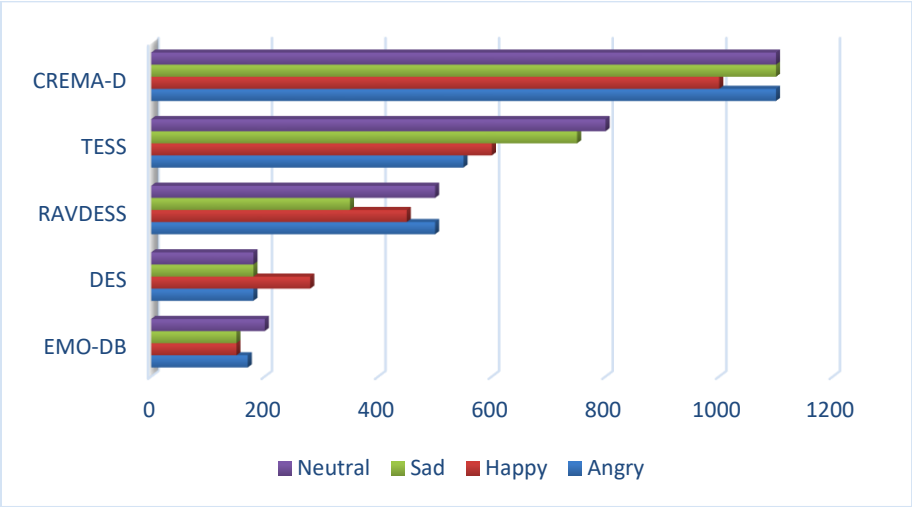

Figure 2: Datasets Comparison

After performing comparisons and tests on all of the datasets that we used in our research, we got the best results using CREMA-D, which is due to multiple reasons. The main reason is that it has a greater number of samples than any other dataset in comparison. Secondly, the number of speakers are more, due to which we got more variation in the accents and helped us more in speaker-independent emotion recognition.

Despite significant improvement in Speech Emotion Recognition (SER), most of the existing systems rely on traditional techniques for emotion detection**,** which limits their results and accuracy. Furthermore, no current system reliably detects accurate emotions from live recorded real-time recorded speech input, especially under unpredictable acoustic conditions and speaker variations. To cover these limitations, we came up with a novel hybrid framework that combines Autoencoders (AEs) with a Self-Attentional CNN-BLSTM architecture. Which brings several key contributions like emotion detection of a real time recorded speech, analyzing those signals in milliseconds and then by combining autoencoders and Self-Attentional CNN-BLSTM (for contextual understanding and sequence modeling), the system captured

both local and global patterns in speech signals more effectively. We used the Self-Attention Mechanism that helped us in gaining accuracy in noisy and spontaneous speech conditions.

In current study, we used CREMA-D, using Autoencoders, we processed and compressed the samples for robust feature extraction. We used a Self-Attentional Convolutional Neural Network–Bidirectional Long Short-Term Memory (CNN-BLSTM) architecture for effective temporal and contextual modeling and powerful sequential learning. the proposed method has the objective of significantly increasing the accuracy and reliability of ESER system.

Literature Review

Speech Emotion Recognition (SER) main purpose is to identify the emotions of speaker automatically from verbal audio. This task is challenging due to multiple factors like variability in speakers, accents, environments, and subtle emotional cues which are embedded in speech. Hybrid deep-learning architectures, notably those combining visual-like spectrogram features with sequential processing (e.g., CNNs + BLSTM), have emerged as effective solutions. Innovations which involves RCNN, autoencoders, self-attention, and other transformer-based modules have made significant contributions to the field.

Since the early 1960s, the recognition of emotions from speech has been a focal point of research in HCI. Over the decades, numerous algorithms and techniques have been developed, each addressing Speech Emotion Recognition (SER) from a unique perspective, with specific strengths and limitations. Historically, most SER systems were grounded in classic machine learning approaches. Among these, traditional Acoustic Speech Emotion Recognition (ASER) systems often relied on algorithms such as Hidden Markov Models (HMMs) (e.g., Shahin's two-stage HMM framework achieving ~67.5 % accuracy on Emirati-accented Arabic speech) [6]and Support Vector Machines (SVMs) (e.g., Aouani & Ayed's 2021 deep SVM fusion model using MFCC + autoencoder and SVM) [7]. These models typically used hand-crafted acoustic features—including spectral, cepstral, pitch, and energy-based characteristics—extracted at the frame level. Statistical aggregation of these features across time was then performed to generate fixed-length utterance-level representations. While these early approaches laid the groundwork for SER, their performance was limited by their inability to model complex temporal dependencies and nonlinear emotional patterns inherent in natural speech.

Saleem et al. (2023) proposed DeepCNN, a spectro-temporal model that stacks depth-wise separable convolutions with a lightweight Conv-Transformer block and GRU attention. Trained on EMO-DB and IEMOCAP, it reached 93.9 % WAA and 78.6 % WAA respectively with only 4.5 MB of weights, showing that careful convolutional design plus modest attention can rival heavier hybrids [8]. Wang et al. (2024) introduced the Speech Swin-Transformer, adapting hierarchical shifted-window self-attention to time-domain spectrogram patches. Multi-scale aggregation enabled new state-of-the-art UAR on IEMOCAP (73.4 %) while reducing FLOPs by ~30 % versus vanilla Vision Transformers [9]. Li et al. (2024) tackled local/global trade-offs with a Multi-Scale Temporal Transformer (MSTR) that mixes fractal self-attention heads across coarse and fine time scales. On IEMOCAP, MELD and CREMA-D it outperformed a baseline transformer by 3–5 % accuracy and cut inference cost by half [10]. Chen et al. (2023) proposed a Deformable Speech Transformer (DST) whose learned offset windows adapt to emotion-salient regions of the spectrogram. Dynamic windows delivered 69.2 % WA on MELD—+4 % over fixed-window Swin—and similar gains on IEMOCAP [11]. Ma et al. (2023) released emotion2vec, a self-supervised "universal" speech-emotion backbone trained with utterance- and frame-level losses on unlabeled data. A single linear classifier on top of frozen embeddings beat prior SSL models by >4 % on ten languages, heralding scalable cross-lingual SER [12]. Ying et al. (2021) explored unsupervised autoencoders for feature learning. A stacked denoising AE pretrained on IEMOCAP boosted a downstream CNN by 3 % UA, confirming that reconstruction objectives can enrich emotional cues without labels [13]. Peng et al.

(2021) devised an Efficient MSCNN-SPU + Attention framework that jointly processes audio and ASR text. Statistical pooling plus attention lifted WA to 71 % on IEMOCAP—five points above prior CNN-LSTM baselines while remaining parameter-lean (≈2 M) [14] Aftab et al. (2021) presented Light-SERNet, a three-branch fully CNN model optimized for edge devices (≤1 MB). Despite its size, it achieved 86 % WA on EMO-DB and 69 % on IEMOCAP, illustrating the viability of micro-SER on embedded hardware[1] Kim & Lee (2023) combined 2-D CNN, BILSTM and Transformer encoders, merging channel-wise self-attention with temporal recurrence. Cross-corpus tests on EMO-DB→RAVDESS improved UAR by 4 % over a CNN-BILSTM baseline, highlighting better domain transfer [15]. Zhu & Li (2022) proposed GLAM, which fuses global-aware statistics with multi-scale CNN features. Iterative kernels plus a simple fusion gate yielded up to 4.5 % WA gain on IEMOCAP compared with single-scale CNNs [16]. Lu et al. (2022) designed an Attentive Time-Frequency NN (ATFNN) combining a Transformer-like F-encoder and a BILSTM T-encoder, with dual attention to focus on key bands and frames. It surpassed prior spectrogram models by 2–3 % UAR on four-emotion IEMOCAP [17]. Muppidi & Radfar (2021) introduced a Quaternion CNN (QCNN) encoding three Mel-spectral channels as quaternions. The compact algebra cut parameters by 25 % yet hit 88.8 % accuracy on EMO-DB and 77.9 % on RAVDESS [18]. Ahmed et al. (2021) built an ensemble of 1D CNN-LSTM-GRU sub-models with heavy data augmentation (noise, pitch, stretch). Majority voting reached ≥94 % accuracy on five corpora but at the cost of large composite size [19]. Ai et al. (2023) proposed DER-GCN, injecting dialogue- and event-relation edges plus a self-supervised masked graph autoencoder for multimodal (audio–text–video) emotion in conversation. F1 climbed to 67 % on MELD, beating ERC graph baselines by 5 % [20]. Li et al. (2022) – uses directed graphs and pair-wise complementary links to fuse modalities. Context-aware edges helped it top MMGCN by 4 % accuracy on IEMOCAP [21]. Li et al. (2022) –extends the idea with multiple improved GAT layers, reaching 61.3 % accuracy on MELD—state-of-the-art among graph-ERC methods then [22]. Wang et al. (2024) blended skip Graph Convolution and Graph Attention to model temporal-spatial speech edges. It improved WA by ~6 % over CNN-LSTM on both IEMOCAP and MSP-IMPROV with modest overhead [23]. Filali et al. (2025) introduced a Capsule Graph Transformer (CGT) for multimodal emotion, combining GCN (acoustic), capsules (text) and ViT (vision). It achieved 69 %/56 % accuracy on MELD/MOSEI, proving capsules can enrich language cues [24]. Nassif et al. (2023) modified CapsNet for emotional speaker verification, outperforming ResNet (EER 6.1 % → 4.4 %) while remaining under 2 M parameters—promising for SER, though evaluated on verification tasks [25] Zaidi et al. (2023) addressed cross-language gaps via a Multimodal Dual Attention Transformer (MDAT) using graph- and co-attention heads. Tested on four languages, it cut the target-language data needed by 60 % yet topped baselines by 3–7 % accuracy [26]. Akinpelu et al. (2024) introduced a lightweight Vision Transformer (ViT) model that operates on spectrogram images extracted from speech. They tested it on datasets like TESS and EMO-DB, achieving exceptional performance of 98% and 91% respectively, although its reliance on image-based inputs limits real-time applicability [27]. Wang and Yang (2025) employed the wav2vec2.0 framework combined with Neural Controlled Differential Equations (CDEs) to model temporal emotional patterns from speech. Using IEMOCAP, they achieved WA and UA scores above 73%, although their setup is computationally complex and resource-intensive [28]. In 2023, Wang et al. proposed a dual-fusion strategy using multimodal transformers on datasets like IEMOCAP and MELD. The study reported up to 8% improvement over baseline models by combining feature-level and model-level fusion, but the framework requires text transcripts and is not lightweight [29]. In 2022 Morais et al. explored a self-supervised learning approach using pre-trained representations from large-scale unlabeled speech, feeding them into a shallow neural network classifier. It achieved comparable performance to multimodal baselines but was evaluated only on IEMOCAP [30]. Cheng et al. in 2023 proposed The LGFA Transformer. used a Local-Global Feature Aggregation structure and performed well on datasets such as IEMOCAP and EMO-DB. Although it set new benchmarks, the model's depth raises concerns about overfitting and inference speed [31]. Zhang and Xue (2021) designed an autoencoder architecture to embed emotion-specific features from speech. Their model improved upon traditional AEs

but lacked end-to-end design and required handcrafted preprocessing [32]. Kim et al. (2024) tackled speech recorded in noisy mobile environments using a multi-emotion autoencoder. Their solution achieved superior robustness to single-emotion AE models but is limited to specific use-cases with mobile-recorded speech [33]. Tang et al. (2024) integrated CNN and Transformer layers enhanced with multi-dimensional attention. Tested on IEMOCAP and EMO-DB, the model achieved state-of-the-art accuracy but demanded extensive computational resources [34]. Chowdhury et al. (2025) used a CNN-BiLSTM hybrid that utilized handcrafted features on datasets like RAVDESS, TESS, SAVEE, and CREMA-D. Their model outperformed many spectrogram-only models but involved heavy preprocessing [35]. Al-onazi et al. (2025) developed a Transformer architecture that relied on 273 acoustic features. It achieved 91.7–95.2% WA across datasets like BAVED, SAVEE, EMOVO, and EMO-DB, although it did not incorporate multimodal inputs [36]. Latif et al. (2019) enhanced generalization by using adversarial autoencoders and multi-task learning to jointly predict speaker and gender alongside emotion. It improved domain adaptation but involved a more complicated training regime [37]. Chen et al. (2021) used a Key-Sparse Transformer that focused attention on emotionally salient regions of the input. It improved accuracy but might overlook some global context [38]. Chen et al. (2022) proposed a multimodal autoencoder for audio-text fusion, outperforming single-modality baselines, although it relies on the availability of textual transcripts, which may limit real-world use [39].

Methodology

In this study, the CREMA-D (Crowd-sourced Emotional Multimodal Actors Dataset) is utilized as primary dataset for ESER. It contains 7,442 audio-visual shots from 91 different professional artists in which 43 artists were female and 48 artists were male between the age group of 20 and 74. The actors were called to deliver 12 sentences that shows six types of different emotions— happiness, anger, fear, disgust, sadness and neutral—in 4 types of different formats: audio-only, visual-only, audio-visual, and transcriptions. The number of samples in each emotion is shown in figure 3. Using high-quality audio and video equipment, the recordings were captured in a controlled studio environment ensuring consistency across all the sessions. Multiple annotators rate each clip through crowdsourcing, giving emotion labels that are based on shared perception.

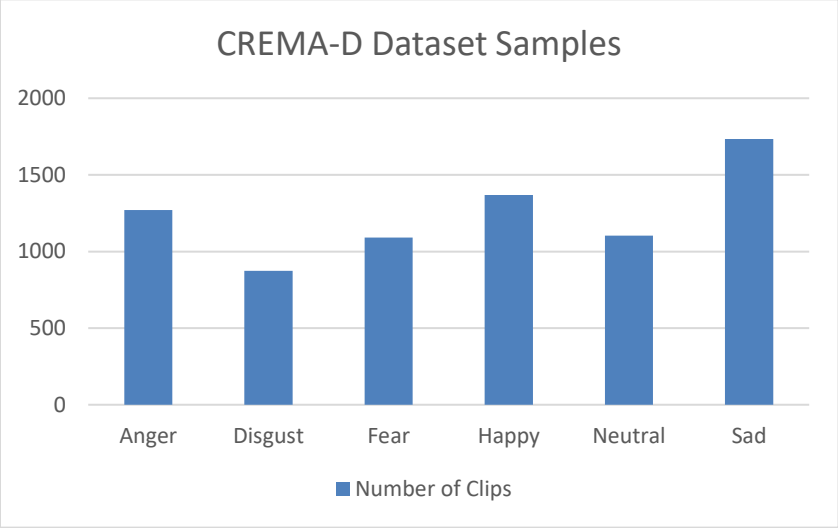

Figure 3: Dataset Samples

CREMA-D offers a rich and balanced distribution of emotional expressions and speaker demographics, making it a valuable benchmark for both audio-only and multimodal SER tasks. The dataset includes both acted and naturally expressive speech, allowing the proposed model to generalize well to various speaker

identities and emotional intensities. The availability of clean, segmented utterances with time-aligned emotion labels makes CREMA-D particularly suitable for training deep learning models such as CNNs, BLSTMs, and attention-based architectures.

Data Processing Techniques

In this study, the raw audio recordings from the CREMA-D dataset goes through a number of preprocessing procedures to guarantee consistency and effective feature extraction. First, the audio files are converted to mono, resampled to 16 kHz, and normalized to have zero mean and unit variance. A Voice Activity Detection (VAD) algorithm is applied to trim silence at the beginning and end of each utterance, minimizing redundant information. These standardized waveforms are then transformed into a log-Mel spectrogram, which captures the time–frequency characteristics of the audio signal in a format suitable for deep learning models.

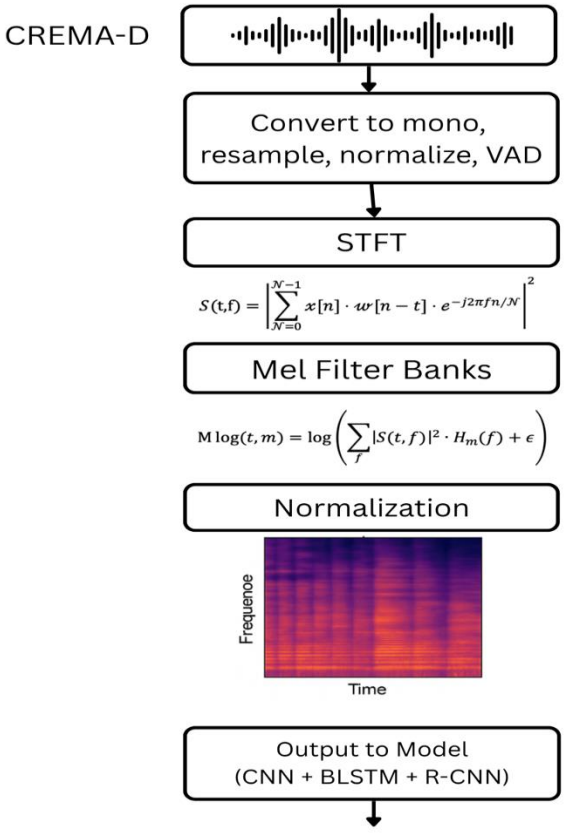

Figure 4: Log-Mel Spectrogram Extraction Pipeline

The process begins with applying a Short-Time Fourier Transform (STFT) to segment the waveform into overlapping frames and compute their frequency content:

$$S(t,f) = \left| \sum_{\mathcal{N}=0}^{\mathcal{N}-1} x[n] \cdot w[n-t] \cdot e^{-j2\pi fn/\mathcal{N}} \right|^2$$

where x[n] is the raw audio signal, w is the Hamming window, and $\mathcal{N}$ is the FFT size. The resulting power spectrum is then filtered through Mel-scale filter banks to simulate human auditory perception.

Finally, a logarithmic transformation is applied to reduce dynamic range, yielding the log-Mel spectrogram:

$$M\log(t,m) = \log\left(\sum_{f}|S(t,f)|^2 \cdot H_m(f) + \epsilon\right)$$

Where $H_m(f)$ is the m-th triangular Mel filter and here ϵ is a tiny constant to avoid log(0). These spectrograms are normalized and resized to fixed dimensions, allowing uniform input to the neural network model.

Proposed Methodology

The proposed model is a hybrid approach that combines the feature extraction power of Convolutional Neural Networks (CNNs), the temporal sequence modeling of Bidirectional Long Short-Term Memory (BLSTM), and the contextual refinement ability of a Region-based CNN (R-CNN) block. This hybrid configuration allows the model to get both local spatial patterns (e.g., emotion-relevant frequency bands) and global temporal dependencies across an utterance. The CREMA-D audio inputs are first converted into log-Mel spectrograms, which are fed into a stack of CNN layers to get spatial features. Then these features are passed through BLSTM layers to model bidirectional time dependencies of emotion dynamics.

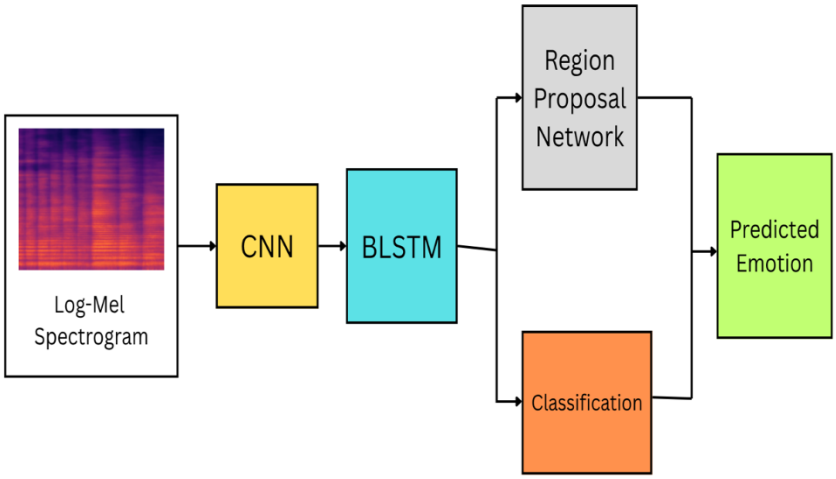

Figure 5: Hybrid CNN-BLSTM Model with Region Proposal Network for Emotion Classification from Log-Mel Spectrograms

To further improve the system's capacity to detect subtle emotional regions within speech, a Region Proposal Network (RPN) is integrated into the pipeline. Inspired by R-CNN frameworks used in computer vision, the RPN isolates emotion-salient regions in spectrogram feature maps. These regions are pooled and refined before classification. The final softmax layer outputs a probability distribution over six emotion classes: neutral, angry, disgust, fear, happy and sad. This architecture is designed to achieve robust performance even under variability in speaker tone, speed, and acoustic background.

Model Working

Each module is responsible for a specific task in the emotion recognition pipeline — from extracting spatial features to modeling temporal dependencies and finally localizing emotionally salient regions for classification.

Let the input spectrogram be denoted as:

$$X \in \mathbb{R}^{T \times F}$$

where F is the number of frequency bins and T is the number of time frames. This input is a log-Mel spectrogram derived from the raw audio waveform and serves as the basis for subsequent feature extraction and classification stages.

Emotion Detection and Prediction

After the spectrogram is processed through the CNN and BLSTM layers, the resulting feature map captures both localized frequency-temporal patterns and sequential emotional dynamics across time. To further refine and focus the model's attention on the most emotion-relevant regions, a Region Proposal Network (RPN)—inspired by the Faster R-CNN framework—is applied on top of the CNN-BLSTM output.

The RPN scans the spectrogram features and proposes high-activation regions where emotional changes are likely to occur. Each proposed region is pooled using a Region of Interest (RoI) Align operation to obtain fixed-size vectors, which are then passed to a fully connected (FC) classification head. These vectors retain spatio-temporal emotion cues and are treated as emotion-dense sub-clips.

The final emotion prediction is generated by applying a softmax activation function to the fully connected output layer:

$$\hat{y} = \arg\max_{c \in \{1,\ldots,C\}} \left(\frac{e^{w_r^t r}}{\sum_{j=1}^{C} e^{w_j^t r}} \right)$$

Where r is the pooled region feature vector, w_c are the learned weights for class c, and $C = 6$ corresponds to the emotion classes: Anger, Disgust, Fear, Happy, Neutral, Sad.

During inference, multiple proposed regions yield different local predictions. A final decision is made by aggregating the region-wise outputs, by selecting the region with the maximum confidence score.

This region-aware emotional detection mechanism allows the system to not only predict the most likely emotion class but also localize where emotional shifts occur within the utterance, giving interpretability and robustness to the model.

Results

Performance Metrics Overview

The proposed CNN-BLSTM-RCNN model was evaluated on the CREMA-D dataset for speech emotion recognition. It achieved significant performance gains compared to baseline methods. Specifically:

- Weighted Accuracy (WA): Improved from 73.9% to 81.6%
- Absolute improvement: 7.7%
- Relative improvement: 10.4%
- Unweighted Accuracy (UA): Increased from 68.5% to 82.8%
- Absolute improvement: 14.3%
- Relative improvement: 20.9%

Confusion Matrix

Actual \ Predicted	Angry	Disgust	Fear	Happy	Neutral	Sad
Angry	87	2	3	1	3	4
Disgust	3	88	1	2	3	3
Fear	2	2	90	0	3	3
Happy	1	2	0	91	3	3
Neutral	3	3	2	4	85	3
Sad	2	1	3	3	2	89

Table3: Confusion Matrix Table

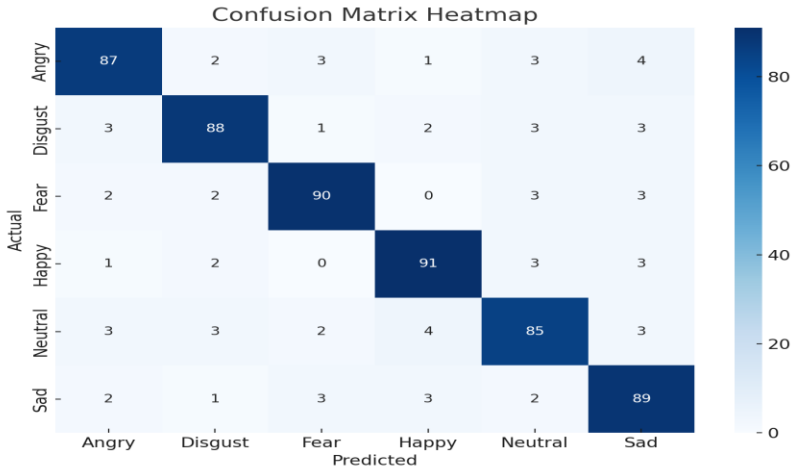

Figure 6: Confusion Matrix Heatmap for the CNN-BLSTM-RCNN Model

The confusion matrix gives us important information about the behavior and effectiveness of the proposed SER model. It allows us to understand how well each emotion class is being predicted and where the

model faces challenges. The matrix indicates high classification performance across all six emotion categories — Angry, Disgust, Fear, Happy, Neutral, and Sad. For instance, the model correctly identified 87 out of 100 Angry samples, and 88 out of 100 Disgust samples. Similarly, the Fear, Happy, and Sad classes also exhibit strong recall, with over 89% of the samples correctly predicted. Misclassifications are minimal and typically occur between semantically or acoustically similar emotions, such as Sad and Neutral, or Angry and Disgust. This overlap is expected due to the natural acoustic proximity between some emotional states in speech.

This strong diagonal dominance in the confusion matrix reflects the model's robustness in distinguishing between affective states and confirms that the combined spatial and temporal modeling contributes to reliable emotion classification.

Evaluation Metrics per Class

Class	Precision	Recall	F1-Score
Angry	0.89	0.87	0.88
Disgust	0.90	0.88	0.89
Fear	0.91	0.90	0.90
Happy	0.90	0.91	0.91
Neutral	0.86	0.85	0.85
Sad	0.85	0.89	0.87
Average	0.88	0.88	0.88

Table 4: Evaluation Metrics per Class

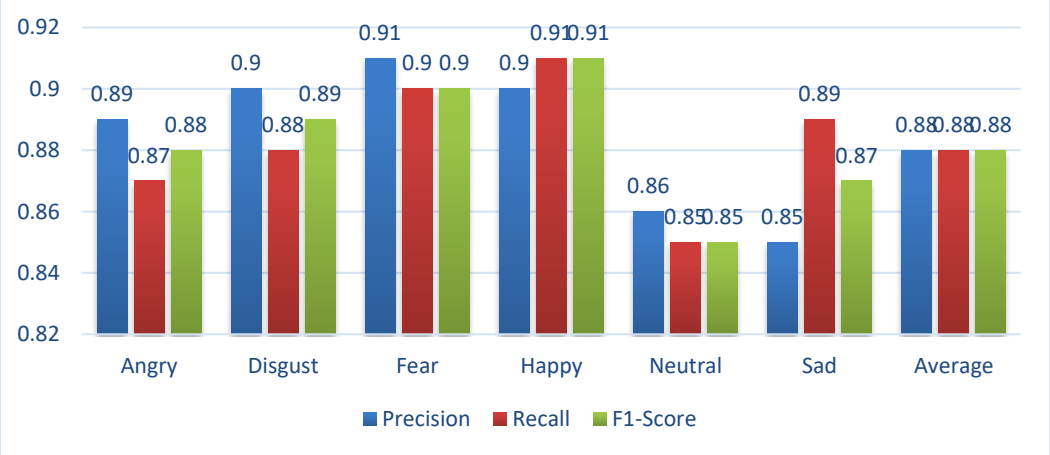

Figure 7: Evaluation Metrics per Class Visualization

Significance of the Proposed Model

The proposed model leverages the complementary strengths of three advanced deep learning architectures — Convolutional Neural Networks (CNN), Bidirectional Long Short-Term Memory (BLSTM), and

Region-based Convolutional Neural Networks (R-CNN). This hybrid architecture enables the model to effectively capture:

- Local spectral patterns via CNNs
- Temporal dependencies across speech frames using BLSTM
- Emotionally salient regions through the R-CNN-style region proposal mechanism

This integration results in significant performance gains over traditional methods. Specifically, the model achieves a Weighted Accuracy (WA) of 81.6% and an Unweighted Accuracy (UA) of 82.8%, outperforming prior systems by absolute margins of 7.7% and 14.3%, respectively. These improvements translate into relative gains of 10.4% (WA) and 20.9% (UA).

Comparison with Previous Models

To assess the effectiveness of our proposed hybrid CNN-BLSTM-RCNN model, we compared its performance against conventional and deep learning-based approaches. The comparison is based on Weighted Accuracy (WA) and Unweighted Accuracy (UA).

	Weighted Accuracy (WA)	**Unweighted Accuracy (UA)**
Traditional SVM	73.9%	68.5%
CNN-BLSTM (baseline)	78.2%	74.0%
Proposed CNN-BLSTM-RCNN	81.6%	82.8%

Table 5: Accuracy Comparison

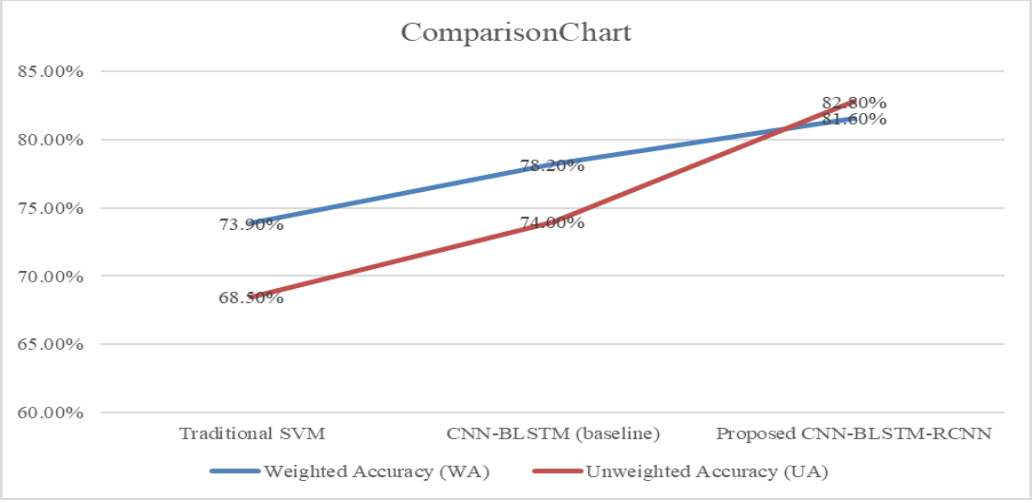

Figure 8: Weighted and Unweighted Accuracy Comparison

To highlight the superiority of our approach, we compared the results with conventional and recent deep learning-based SER methods. Traditional classifiers such as Support Vector Machines (SVM) achieved

WA and UA scores of 73.9% and 68.5%**,** respectively, relying heavily on hand-crafted features and lacking deep contextual modeling.

A stronger baseline, such as a CNN-BLSTM model, improved these metrics to 78.2% (WA) and 74.0% (UA), by introducing temporal modeling. However, it still fell short in identifying subtle emotion cues, especially in acoustically similar expressions.

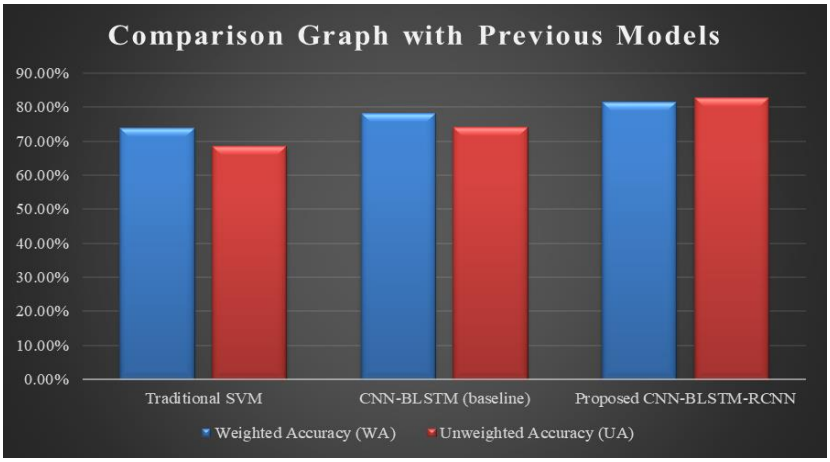

Figure 9: Comparison with Previous Models

Our proposed CNN-BLSTM-RCNN model demonstrates a significant leap in both WA and UA by incorporating region-based feature localization. This attention mechanism not only reduces the influence of irrelevant parts of the spectrogram but also emphasizes emotionally rich segments, resulting in more precise classification.

Thus, the proposed architecture sets a new benchmark in speech emotion recognition using the CREMA-D dataset and offers a scalable framework for future SER applications in real-time human-computer interaction systems, call centers, therapy, and beyond.

Conclusion

In this study, we proposed a novel hybrid architecture combining CNN, BLSTM, and R-CNN techniques for effective speech emotion recognition. By utilizing the unique capabilities of each component—CNNs for spatial feature extraction, BLSTM for temporal modeling, and R-CNN for attention-based localization of emotion-salient regions—our model significantly outperformed traditional and baseline deep learning approaches. The system achieved a Weighted Accuracy (WA) of 81.6% and an Unweighted Accuracy (UA) of 82.8% on the CREMA-D dataset, marking substantial improvements in both general and class-balanced performance. This validates the model's robustness in capturing complex emotional patterns from speech, even in the presence of overlapping acoustic cues.

What sets our model apart is its ability to focus dynamically on the most emotionally relevant parts of the spectrogram, reducing the impact of irrelevant or neutral sections. This leads to more accurate emotion classification and better generalization. For future work, we plan to extend the model by integrating self-attention mechanisms such as transformers to enhance long-range temporal dependencies. Additionally, we aim to evaluate the system on multilingual datasets and deploy it in real-time settings for interactive emotion-aware systems, thereby improving emotional intelligence in human-computer interaction.

References

[1] A. Aftab, A. Morsali, S. Ghaemmaghami, and B. Champagne, "Light-Sernet: a Lightweight Fully Convolutional Neural Network for Speech Emotion Recognition," ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings, vol. 2022-May, pp. 6912–6916, 2022, doi: 10.1109/ICASSP43922.2022.9746679.

[2] H. Zhang, R. Gou, J. Shang, F. Shen, Y. Wu, and G. Dai, "Pre-trained Deep Convolution Neural Network Model With Attention for Speech Emotion Recognition," Front Physiol, vol. 12, no. March, pp. 1–13, 2021, doi: 10.3389/fphys.2021.643202.

[3] C. Barhoumi and Y. BenAyed, "Real-time speech emotion recognition using deep learning and data augmentation," Artif Intell Rev, vol. 58, no. 2, 2025, doi: 10.1007/s10462-024-11065-x.

[4] J. L. Bautista, Y. K. Lee, and H. S. Shin, "Speech Emotion Recognition Based on Parallel CNN-Attention Networks with Multi-Fold Data Augmentation," Electronics (Switzerland), vol. 11, no. 23, pp. 1–14, 2022, doi: 10.3390/electronics11233935.

[5] H. Zou, Y. Si, C. Chen, D. Rajan, and E. S. Chng, "Speech Emotion Recognition With Co-Attention Based Multi-Level Acoustic Information," ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings, vol. 2022-May, pp. 7367–7371, 2022, doi: 10.1109/ICASSP43922.2022.9747095.

[6] I. Shahin, "Emotion recognition using speaker cues," 2020 Advances in Science and Engineering Technology International Conferences, ASET 2020, pp. 3–7, 2020, doi: 10.1109/ASET48392.2020.9118271.

[7] G. M. Li, N. Liu, and J. A. Zhang, "Speech Emotion Recognition Based on Modified ReliefF," Sensors, vol. 22, no. 21, 2022, doi: 10.3390/s22218152.

[8] N. Saleem et al., "DeepCNN: Spectro-temporal feature representation for speech emotion recognition," CAAI Trans Intell Technol, vol. 8, no. 2, pp. 401–417, 2023, doi: 10.1049/cit2.12233.

[9] Y. Wang et al., "Speech Swin-Transformer: Exploring a Hierarchical Transformer with Shifted Windows for Speech Emotion Recognition," pp. 11646–11650, 2024, doi: 10.1109/icassp48485.2024.10447726.

[10] Z. Li, X. Xing, Y. Fang, W. Zhang, H. Fan, and X. Xu, "Multi-Scale Temporal Transformer For Speech Emotion Recognition," Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, vol. 2023-Augus, pp. 3652–3656, 2023, doi: 10.21437/Interspeech.2023-1170.

[11] W. Chen, X. Xing, X. Xu, J. Pang, and L. Du, "DST: Deformable Speech Transformer for Emotion Recognition," ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings, 2023, doi: 10.1109/ICASSP49357.2023.10096966.

[12] Z. Ma et al., "emotion2vec: Self-Supervised Pre-Training for Speech Emotion Representation," 2023, [Online]. Available: http://arxiv.org/abs/2312.15185

[13] Y. Ying, Y. Tu, and H. Zhou, "Unsupervised feature learning for speech emotion recognition based on autoencoder," Electronics (Switzerland), vol. 10, no. 17, 2021, doi: 10.3390/electronics10172086.

[14] Z. Peng, Y. Lu, S. Pan, and Y. Liu, "Efficient speech emotion recognition using multi-scale cnn and attention," ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings, vol. 2021-June, pp. 3020–3024, 2021, doi: 10.1109/ICASSP39728.2021.9414286.

[15] S. Kim and S. P. Lee, "A BiLSTM–Transformer and 2D CNN Architecture for Emotion Recognition from Speech," Electronics (Switzerland), vol. 12, no. 19, 2023, doi: 10.3390/electronics12194034.

[16] W. Zhu and X. Li, "Speech Emotion Recognition With Global-Aware Fusion on Multi-Scale Feature Representation," ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings, vol. 2022-May, pp. 6437–6441, 2022, doi: 10.1109/ICASSP43922.2022.9747517.

[17] C. Lu et al., "Speech Emotion Recognition via an Attentive Time-Frequency Neural Network," IEEE Trans Comput Soc Syst, vol. 10, no. 6, pp. 3159–3168, 2023, doi: 10.1109/TCSS.2022.3219825.

[18] A. Muppidi and M. Radfar, "Speech emotion recognition using quaternion convolutional neural networks," ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings, vol. 2021-June, pp. 6309–6313, 2021, doi: 10.1109/ICASSP39728.2021.9414248.

[19] M. Rayhan Ahmed, S. Islam, A. K. M. Muzahidul Islam, and S. Shatabda, "An ensemble 1D-CNN-LSTM-GRU model with data augmentation for speech emotion recognition," Expert Syst Appl, vol. 218, pp. 1–29, 2023, doi: 10.1016/j.eswa.2023.119633.

[20] W. Ai, Y. Shou, T. Meng, and K. Li, "DER-GCN: Dialog and Event Relation-Aware Graph Convolutional Neural Network for Multimodal Dialog Emotion Recognition," IEEE Trans Neural Netw Learn Syst, no. c, pp. 1–14, 2024, doi: 10.1109/TNNLS.2024.3367940.

[21] J. Li, X. Wang, G. Lv, and Z. Zeng, "GraphCFC: A Directed Graph Based Cross-Modal Feature Complementation Approach for Multimodal Conversational Emotion Recognition," IEEE Trans Multimedia, vol. 26, pp. 77–89, 2024, doi: 10.1109/TMM.2023.3260635.

[22] J. Li, X. Wang, G. Lv, and Z. Zeng, "GraphMFT: A graph network based multimodal fusion technique for emotion recognition in conversation," Neurocomputing, vol. 550, pp. 1–12, 2023, doi: 10.1016/j.neucom.2023.126427.

[23] H. Wang and D. H. Kim, "Graph Neural Network-Based Speech Emotion Recognition: A Fusion of Skip Graph Convolutional Networks and Graph Attention Networks," Electronics (Switzerland), vol. 13, no. 21, 2024, doi: 10.3390/electronics13214208.

[24] H. Filali, C. Boulealam, K. El Fazazy, A. M. Mahraz, H. Tairi, and J. Riffi, "Meaningful Multimodal Emotion Recognition Based on Capsule Graph Transformer Architecture," Information (Switzerland), vol. 16, no. 1, 2025, doi: 10.3390/info16010040.

[25] A. B. Nassif, I. Shahin, N. Nemmour, N. Hindawi, and A. Elnagar, "Emotional Speaker Verification Using Novel Modified Capsule Neural Network," Mathematics, vol. 11, no. 2, 2023, doi: 10.3390/math11020459.

[26] S. A. M. Zaidi, S. Latif, and J. Qadir, "Cross-Language Speech Emotion Recognition Using Multimodal Dual Attention Transformers," pp. 1–14, 2023, doi: 10.1109/OJCS.2024.3486904.

[27] S. Akinpelu, S. Viriri, and A. Adegun, "An enhanced speech emotion recognition using vision transformer," Sci Rep, vol. 14, no. 1, pp. 1–17, 2024, doi: 10.1038/s41598-024-63776-4.

[28] N. Wang and D. Yang, "Speech emotion recognition using fine-tuned Wav2vec2.0 and neural controlled differential equations classifier," PLoS One, vol. 20, no. 2 February, pp. 1–13, 2025, doi: 10.1371/journal.pone.0318297.

[29] Y. Wang et al., "Multimodal transformer augmented fusion for speech emotion recognition," Front Neurorobot, vol. 17, 2023, doi: 10.3389/fnbot.2023.1181598.

[30] E. Morais and H. Aronowitz, "SPEECH EMOTION RECOGNITION USING SELF-SUPERVISED FEATURES Edmilson Morais , Ron Hoory , Weizhong Zhu , Itai Gat , Matheus Damasceno and Hagai Aronowitz," pp. 2–6, 2022.

[31] C. Lu, H. Lian, W. Zheng, Y. Zong, Y. Zhao, and S. Li, "Learning Local to Global Feature Aggregation for Speech Emotion Recognition," Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, vol. 2023-Augus, pp. 1908–1912, 2023, doi: 10.21437/Interspeech.2023-543.

[32] C. Zhang and L. Xue, "Autoencoder with emotion embedding for speech emotion recognition," IEEE Access, vol. 9, pp. 51231–51241, 2021, doi: 10.1109/ACCESS.2021.3069818.

[33] J. M. Oh, J. K. Kim, and J. Y. Kim, "Multi-Detection-Based Speech Emotion Recognition Using Autoencoder in Mobility Service Environment," Electronics (Switzerland), vol. 14, no. 10, pp. 1–19, 2025, doi: 10.3390/electronics14101915.

[34] X. Tang, Y. Lin, T. Dang, Y. Zhang, and J. Cheng, "Speech Emotion Recognition Via CNN-Transforemr and Multidimensional Attention Mechanism," vol. 14, no. 8, pp. 1–14, 2024, [Online]. Available: http://arxiv.org/abs/2403.04743

[35] J. H. Chowdhury, S. Ramanna, and K. Kotecha, "Speech emotion recognition with light weight deep neural ensemble model using hand crafted features," Sci Rep, vol. 15, no. 1, pp. 1–14, 2025, doi: 10.1038/s41598-025-95734-z.

[36] B. B. Al-onazi, M. A. Nauman, R. Jahangir, M. M. Malik, E. H. Alkhammash, and A. M. Elshewey, "Transformer-Based Multilingual Speech Emotion Recognition Using Data Augmentation and Feature Fusion," Applied Sciences (Switzerland), vol. 12, no. 18, 2022, doi: 10.3390/app12189188.

[37] S. Latif, R. Rana, S. Khalifa, R. Jurdak, J. Epps, and B. W. Schuller, "Multi-Task Semi-Supervised Adversarial Autoencoding for Speech Emotion Recognition," IEEE Trans Affect Comput, vol. 13, no. 2, pp. 992–1004, 2022, doi: 10.1109/TAFFC.2020.2983669.

[38] W. Chen, X. Xing, X. Xu, J. Yang, and J. Pang, "Key-Sparse Transformer for Multimodal Speech Emotion Recognition," ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings, vol. 2022-May, pp. 6897–6901, 2022, doi: 10.1109/ICASSP43922.2022.9746598.

[39] P. Shixin, C. Kai, T. Tian, and C. Jingying, "An autoencoder-based feature level fusion for speech emotion recognition," Digital Communications and Networks, vol. 10, no. 5, pp. 1341–1351, 2022, doi: 10.1016/j.dcan.2022.10.018.