Speech Emotion Recognition using Convolutional-Recurrent Hybrid Neural Networks

Mahesh Divakaran Namboodiri
MS in Computer Engineering
Arizona State University
mnamboo1@asu.edu

Samudyata Sudarshan Jagirdar
MS in Computer Engineering
Arizona State University
sjagird1@asu.edu

Sayantika Paul
MS in Computer Engineering
Arizona State University
spaul56@asu.edu

***Abstract*—Speech Emotion Recognition (SER) is a key technology in improving human-computer interaction, enabling machines to detect emotions conveyed through speech and thus offering richer, more intuitive communication experiences. This paper presents a comprehensive literature review of various SER approaches, ranging from traditional methods to more recent deep learning architectures such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs). A particular focus is placed on hybrid CNN-RNN models, which combine spatial feature extraction and temporal dependency modeling to address the complexities of emotion recognition in speech. Popular datasets, including RAVDESS, TESS, and CREMA-D, are analyzed for their contribution to SER benchmarking, highlighting their role in improving model generalization across diverse emotional states. This review identifies the strengths and limitations of different models, emphasizing challenges such as speaker variability, noise resilience, and the need for real-time processing. Through this analysis, our goal is to provide a foundation for future research, exploring how SER systems can be further enhanced for applications in mental health diagnostics, customer service, and more.**

***Index Terms*—Speech Emotion Recognition (SER), CNN-RNN Hybrid Architecture, Temporal Dependencies, Spatial Features, Deep Learning, Mel-Spectrograms, Spectral Subtraction, RAVDESS, TESS, CREMA-D, Noise Resilience, Human-Computer Interaction (HCI).**

I. Introduction

Emotions are fundamental to human communication, transcending spoken words to convey deeper meanings and intentions. In human-computer interaction (HCI), the ability of machines to recognize emotions from speech can transform how we interact with technology, enabling more intuitive and empathetic responses [1]. Speech Emotion Recognition (SER) involves detecting and classifying emotions in spoken language, with applications spanning from healthcare monitoring to customer service enhancement [3]. While humans naturally excel at perceiving emotional nuances in speech, replicating this capability in machines presents significant challenges due to the complex, non-linear nature of emotional expression in audio signals [17].

The challenge of accurate emotion recognition is compounded by several factors. Speech signals are inherently dynamic, with emotions manifested through variations in pitch, tone, and rhythm that evolve over time [2]. These patterns are further complicated by individual differences in emotional expression, cultural variations, and the presence of background noise in real-world environments [20]. Traditional approaches using linear classifiers have proven inadequate for capturing these nuanced emotional patterns [4], leading researchers to explore more sophisticated methods.

Deep learning has emerged as a promising solution, with Convolutional Neural Networks (CNNs) excelling at spatial feature extraction [9], [10], [11], and Recurrent Neural Networks (RNNs) demonstrating efficacy in modeling temporal patterns [12], [13]. However, existing approaches often focus on either spatial or temporal features independently, limiting their effectiveness in capturing the full spectrum of emotional cues in speech [21]. Additionally, current methods struggle with robustness across different speakers and environmental conditions [18], highlighting the need for more sophisticated approaches.

To address these limitations, this paper presents a novel hybrid CNN-RNN architecture for Speech Emotion Recognition (SER). The contributions of this work are centered on three main aspects:

1) **Advanced Preprocessing Pipeline**: Our approach includes an advanced preprocessing pipeline that incorporates spectral subtraction for noise reduction and feature extraction using Mel-Spectrogram features. This significantly enhances the model's ability to capture emotional variations in speech while maintaining robustness against background noise, which is essential for real-world applications.
2) **Hybrid CNN-RNN Architecture**: We developed an innovative hybrid model that combines Convolutional Neural Networks (CNNs) for spatial feature extraction with Recurrent Neural Networks (RNNs) for temporal pattern recognition. This integration captures both local and

global emotional cues, overcoming the limitations of existing models that treat spatial and temporal features separately.

3) **Comprehensive Evaluation Framework**: We created a robust evaluation framework using multiple datasets (RAVDESS, TESS, and CREMA-D), totaling 10,898 samples from five primary emotions plus neutral states. This framework enables cross-dataset validation, ensuring the model's generalizability across different recording conditions and speaker demographics.

The work in [14] and [15] uses similar architectures but with a focus on only one type of feature extraction. The proposed hybrid CNN-RNN architecture represents a significant advancement in SER technology, offering improved accuracy and robustness compared to existing approaches. By addressing the fundamental challenges of emotion recognition in speech, this work paves the way for more effective applications in healthcare monitoring, customer service, and psychological assessment tools.

II. Related Work

Speech Emotion Recognition (SER) has progressed from traditional handcrafted features to deep learning approaches. The field combines prosodic elements, spectral characteristics, and TEO-based features with various classification methods, from GMM and SVM to CNN+LSTM hybrids, achieving up to 88.6% accuracy for seven emotions [1], [2]. Although CNN + LSTM architectures show promise, achieving 82.35% accuracy on the Berlin database, challenges persist, including limited natural emotion data, cross-lingual issues, and cultural variations in emotion expression [3], [4]. Research uses spontaneous, acted on, and elicited speech corpora, with recent advances focusing on attention mechanisms and multiple classifier methods [5].

The literature on feature selection in speech recognition spans traditional methods like Fuzzy Rough Sets, PSO, and Genetic Algorithms to advanced auditory-inspired approaches [6], [7]. Although conventional techniques such as PCA and LDA are common, newer methods based on human auditory processing show superior performance under noisy conditions [8]. Sequential Feature Selection methods achieve high recognition rates despite computational demands, while spectro-temporal features using clustering (WKM and GMM) demonstrate improved robustness over traditional MFCCs [5]. The research suggests that the combination of multiple approaches may offer optimal performance, balancing computational efficiency with recognition accuracy [8].

Several key datasets have significantly advanced research in speech recognition and emotion detection. For phoneme classification, the **TIMIT** database offers 6,300 sentences from 630 speakers across U.S dialects [6]. In emotion recognition, popular datasets include **IEMOCAP** (12 hours of audiovisual data from 10 actors) [3], **EMO-DB** (Berlin Emotional Database with seven emotion categories) [2], **RAVDESS** [1], **SAVEE** [8], and **SEMAINE** [5]. Specialized datasets like **BHUDES**, **CASIA**, and the **Polish Emotion Speech Database** add cultural diversity [7]. Many studies use multiple datasets for comprehensive evaluation, with **IEMOCAP** and **EMO-DB** being the most cited, achieving accuracy rates from 58.46% to 85.57% [2], [3]. These datasets include acted, spontaneous, and elicited emotions, offering both categorical and dimensional labels for robust model testing.

Convolutional Neural Networks (CNNs) have gained popularity in audio classification tasks, such as speech emotion recognition, by directly extracting emotions from raw speech. A recent study proposed a CNN-based model that achieved 65.7% accuracy on six emotion categories, outperforming conventional SVM methods [16]. CNNs have also been successful in environmental sound classification, as demonstrated by Karol J. Piczak, who used spectrograms for classifying sounds like rain and footsteps [9]. Additionally, CNNs have shown superior performance over traditional DNN models in large vocabulary speech recognition tasks [10].

Recurrent Neural Networks (RNNs) are another widely used model for emotion recognition due to their ability to capture temporal dependencies in sequential data. Mirsamadi et al. [12] introduced an RNN with a local attention mechanism for automatic feature identification in speech, improving on traditional statistical methods. Li et al. [13] enhanced emotion recognition by processing data in both forward and backward directions, outperforming standard RNN and SVM models across various datasets like IEMOCAP and EMO-DB.

To integrate the feature extraction capability of CNNs and the temporal analysis of RNNs, hybrid CNN-RNN models are used. For example, Keunwoo Choi et al. [14] worked on music classification, where the authors use a convolutional recurrent neural network (CRNN), in which CNN layers extract features from audio signals, and RNN layers with gated recurrent units (GRUs) capture temporal information. Another study conducted by Chen et al. [15] on speech emotion recognition introduces a 3D CRNN with an attention mechanism, combining 3D convolutions and LSTM units to improve the recognition of emotions like sadness. These hybrid models outperform traditional CNNs and RNNs in handling complex audio tasks.

Data preprocessing techniques play a crucial role in enhancing Speech Emotion Recognition (SER) performance by addressing challenges like noise and variability in audio signals. One study demonstrates how deep neural networks (DNNs) can model neural adaptation to noise, effectively altering the spectro-temporal receptive field to enable noise filtering

and improve speech perception in dynamic environments [20]. Another approach combines deep learning-based and handcrafted audio features, utilizing models like wav2vec2 and openSmile for feature extraction. Through iterative feature selection and majority voting, this method achieves a 3% improvement in classification accuracy, with a top accuracy of 92.55% on a multi-corpus dataset comprising common emotions like sadness, happiness, and anger [21].

III. Methodology

Our speech emotion recognition model employs a sophisticated pipeline designed to effectively capture both spatial and temporal features of audio data. This approach addresses limitations of previous methods and ensures robust emotion recognition. The pipeline consists of audio input processing, feature extraction, data preparation, model architecture specification, training, and prediction.

A. Audio Data Preprocessing

The preprocessing stage is crucial for ensuring consistency and enhancing the quality of the input data:

- **Resampling**: Audio is resampled to a standard rate of 16 kHz to ensure uniformity across different recordings, facilitating consistent feature extraction.
- **Denoising**: A high-pass filter is applied to remove low-amplitude, high-frequency noise, enhancing the clarity of emotional cues in the audio.
- **Channel Flattening**: Multi-channel audio is converted to mono, reducing computational complexity without significant loss of emotional information.
- **Normalization**: Audio signals are normalized to a fixed amplitude range (-1 to 1) to mitigate discrepancies due to varying recording volumes.
- **Padding/Truncating**: Audio clips are adjusted to a fixed duration (e.g., 5 seconds) by zero-padding or truncating, ensuring consistency for batch processing during training.

This comprehensive preprocessing approach ensures that the subsequent feature extraction stage receives clean, uniform, and information-rich audio inputs.

B. Feature Extraction

The feature extraction phase plays a pivotal role in transforming raw audio signals into a representation suitable for deep learning analysis. This work primarily utilizes mel-spectrogram features, which effectively capture the time-frequency characteristics of speech emotions while considering human auditory perception.

1) **Mel-Spectrogram:** The Mel spectrogram provides a time-frequency representation on a Mel scale:

$$S_{mel}(l,m) = \sum_{k} |X(l,k)|^2 H_m(k) \qquad (1)$$

where $H_m(k)$ are the Mel filter bank responses. This representation is particularly effective for CNN interpretation of audio signals.
The mel-spectrogram features are extracted using the following parameters:

- Number of mel bands (n_mels) = 128
- Fast Fourier Transform (FFT) window size = 2048 samples
- Hop length = 512 samples
- Window length = 2048 samples
- Maximum frequency (f_{max}) = 8000 Hz

Fig. 1 demonstrates the log-scaled spectrogram representation of a random audio sample from the dataset, where the x-axis represents time frames, y-axis shows frequencies (Hz), and the color intensity indicates the magnitude in decibels (dB), with yellow regions representing higher energy concentrations and purple regions showing lower energy areas.

2) **Logarithmic Scaling** To better align with human auditory perception, the power spectrogram is converted to the decibel scale:

$$S_{db} = 10 \log_{10}(S_{mel} + \epsilon) \qquad (2)$$

where ϵ = 1e-10 is added to prevent numerical instability in logarithmic computation.

3) **Feature Normalization** Standardization is applied to ensure consistent scale across all features:

$$S_{norm} = \frac{S_{db} - \mu}{\sigma + \epsilon} \qquad (3)$$

where μ is the mean and σ is the standard deviation of the mel-spectrogram. The normalized features are clipped to the range [-5, 5] to handle extreme values while preserving the majority of the useful signal information.
The final feature representation has dimensions of 128×173×1, where:

- 128 represents the number of mel frequency bands
- 173 represents the time steps
- 1 represents the channel dimension

This diverse set of features ensures a comprehensive representation of the audio signal, capturing both spectral and temporal characteristics relevant to emotion recognition. The mel-scale transformation particularly enhances the model's ability to learn emotion-relevant features, as it approximates human auditory perception.

C. Model Architecture

Our proposed solution implements a CNN-RNN hybrid architecture, effectively combining the strengths of Convolutional Neural Networks (CNNs) for spatial feature extraction and Recurrent Neural Networks

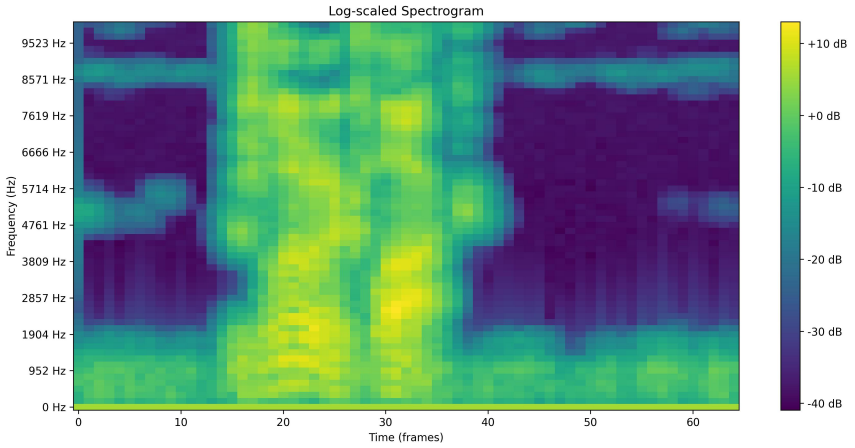

Fig. 1: Spectrogram of a random audio from the dataset

(RNNs) for temporal sequence modeling. This integrated approach enables comprehensive analysis of emotional speech patterns.

The CNN component consists of four convolutional blocks with progressively increasing filters (64, 128, 256, 512). Each block comprises a Conv2D layer with 3×3 kernel and 'same' padding, followed by BatchNormalization, ReLU activation, MaxPooling2D (2×2, stride 2), and Dropout (0.2). This structure efficiently extracts hierarchical spatial features from the mel spectrograms, capturing local patterns and spectro-temporal characteristics.

The RNN component incorporates two LSTM layers: the first with 256 units returning sequences, and the second with 128 units. Each LSTM layer is augmented with BatchNormalization and Dropout (0.3). This configuration enables the modeling of temporal dependencies in the CNN-extracted features, providing a hierarchical representation of temporal information crucial for understanding the evolution of audio features over time.

The dense component consists of two blocks with 256 and 128 units respectively, each comprising Dense layers followed by BatchNormalization, ReLU activation, and Dropout (0.3). The architecture culminates in a final dense layer with softmax activation, where the number of units corresponds to the number of emotion classes. This progressive reduction in units (256 to 128 to num_emotions) enables gradual refinement of features leading to final classification.

D. Training Configuration

The training process is optimized for both performance and generalization through careful parameter selection. We employ the AdamW optimizer with a learning rate of 1×10^{-3} and weight decay of 1×10^{-4}, combined with categorical crossentropy as the loss function. The model is trained with a batch size of 32, incorporating early stopping with a patience of 15 epochs and learning rate reduction with a factor of 0.1, patience of 5, and minimum learning rate of 1×10^{-6}.

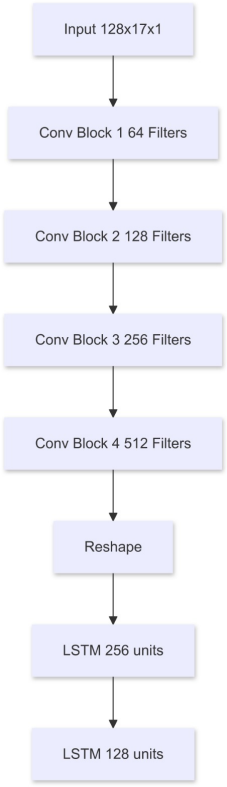

Fig. 2: CNN-RNN Hybrid Architecture

E. Regularization Strategy

To enhance model generalization and prevent overfitting, we implement a comprehensive regularization approach:

- Dropout layers with rates varying from 0.2 to 0.3 throughout the network
- Batch normalization in all major components
- Weight decay implementation in the optimizer
- Early stopping mechanism
- Dynamic learning rate adjustment

This multi-faceted regularization strategy ensures

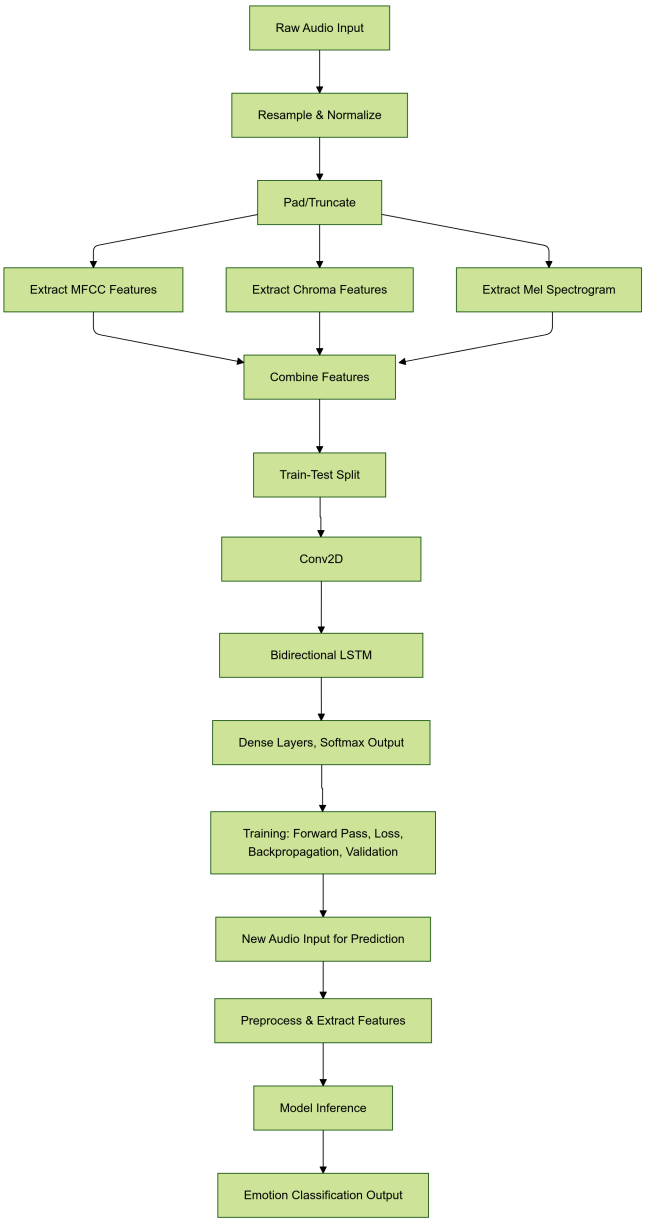

Fig. 3: SER Methodology

robust feature learning while mitigating overfitting risks. The proposed method integrates advanced preprocessing, feature extraction, and a hybrid deep learning architecture to effectively capture the nuances of emotional speech, addressing the inherent complexities in speech emotion recognition tasks.

IV. Dataset

In order to enhance performance results, around 10,000 emotion datasets were combined into a single, unified dataset. This merged dataset includes various emotional categories: Anger, Disgust, Fear, Happy, and Sad. By consolidating these datasets, a more comprehensive and diverse range of emotional data is available, which improves the robustness and accuracy of performance outcomes.

Dataset	Size	Emotions	Type
RAVDESS	1,440 samples	6 emotions	Acted
TESS	2,800 samples	6 emotions	Acted
CREMA-D	7,442 samples	6 emotions	Natural

TABLE I: Datasets Utilized

V. Results and Discussion

A. Model Performance Analysis

The speech emotion recognition model was evaluated on a dataset of 10,898 samples, achieving notable performance metrics:

- Training Accuracy: 98.96%
- Validation Accuracy: 90.26%
- Test Accuracy: 90.26%
- Test Loss: 0.3350

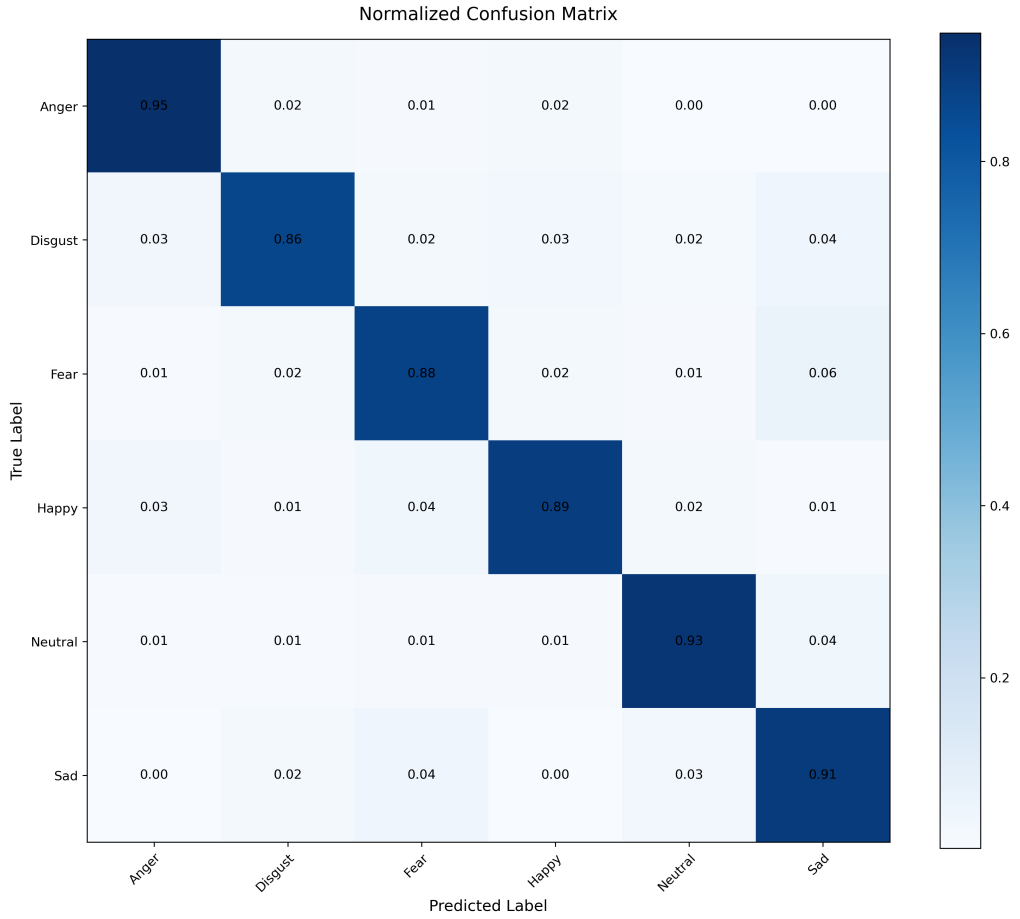

Fig. 4: Normalized confusion matrix showing the model's classification performance across emotion categories

B. Emotion Classification Performance

The model demonstrated robust performance across different emotion categories, with the following F1-scores:

Emotion	F1-Score
Anger	0.93
Disgust	0.89
Fear	0.88
Happy	0.91
Sad	0.88
Neutral	0.92

TABLE II: Per-class F1-scores for emotion classification

C. Detailed Performance Analysis

1) Emotion Classification Precision: The normalized confusion matrix provides nuanced insights into our emotion recognition model's performance (Figure 4):

- **Anger Recognition:** Demonstrates exceptional precision with a 95% accuracy rate, indicating robust and distinctive feature extraction for anger-related emotional states. The model shows remarkable ability to isolate anger from other emotional categories with minimal misclassification.
- **Disgust Classification:** Achieves a solid 86% accuracy, with a minor 4% misclassification rate primarily with sadness. This suggests some subtle overlap in the feature representations of disgust and sad emotional expressions, which could warrant further investigation into distinguishing features.
- **Fear Detection:** Exhibits 88% accuracy, with a 6% confusion rate with sadness. This indicates potential similarities in the underlying facial muscular configurations or feature representations between fear and sad emotional states.
- **Happiness Recognition:** Maintains a high 89% accuracy, demonstrating consistent performance in identifying joyful emotional expressions across diverse facial variations.
- **Neutral Expression Analysis:** Shows exceptional discrimination with a 93% accuracy rate, highlighting the model's proficiency in distinguishing neutral states from other emotional categories.
- **Sadness Classification:** Delivers a robust 91% accuracy, reflecting the model's sophisticated ability to capture the subtle nuances of sad emotional expressions.

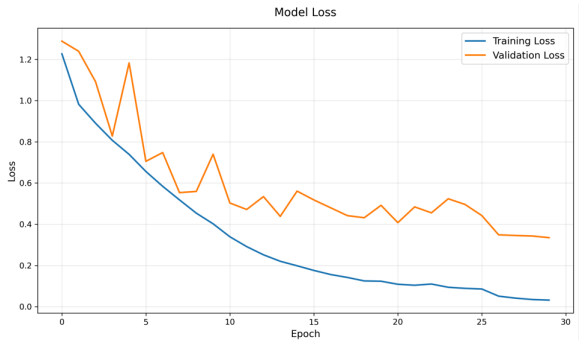

Fig. 5: Training and validation loss curves over epochs

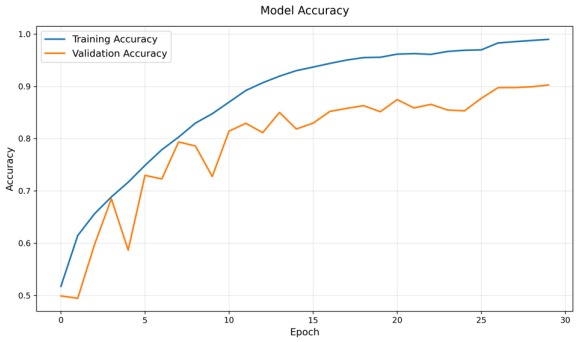

Fig. 6: Training and validation accuracy curves over epochs

2) Model Training Dynamics and Convergence: The training evolution over 30 epochs reveals critical insights into the model's learning characteristics (Figures 5 and 6):

- **Initial Learning Phase:** Characterizd by rapid convergence within the first 5 epochs, indicating the model's quick adaptation to the underlying emotional feature representations.
- **Performance Stabilization:** Validation accuracy plateaus around 90% after epoch 20, suggesting the model has captured the most significant discriminative features for emotion recognition.
- **Loss Trajectory:** A consistent and gradual decrease in training loss throughout the training process, demonstrating the model's progressive refinement of its internal representations.

D. Ablation Study

An ablation study was conducted to systematically evaluate the contribution of key architectural components to the performance of our Speech Emotion Recognition (SER) model. Table 2 highlights the comparative analysis, where the hybrid CNN-RNN model achieved an accuracy of 84.89% on the combined dataset, significantly outperforming the CNN-only and LSTM-only architectures, which achieved 65.7% and 71.2% accuracy on RAVDESS, respectively.

Table III shows our model's performance compared to existing approaches:

Model	Accuracy	Dataset
CNN only	65.7%	RAVDESS
LSTM only	71.2%	RAVDESS
Our Hybrid	84.89%	Combined

TABLE III: Performance Comparison

The CNN-only model excelled in capturing spatial features from spectrograms, such as pitch and intensity, but failed to account for temporal dependencies crucial for understanding emotional transitions in speech. For instance, emotions like sadness or anger often manifest through prolonged or sudden shifts in tone, which static spatial analysis alone cannot discern. Conversely, the LSTM-only model demonstrated improved performance by modeling temporal patterns. However, its inability to extract spatial features effectively limited its capacity to identify localized spectral nuances essential for emotion recognition.

Our hybrid CNN-RNN architecture successfully integrates the spatial feature extraction strengths of CNNs with the temporal sequence modeling capabilities of RNNs. This synergy enables the model to robustly capture the complex and dynamic nature of emotional expressions while maintaining resilience to noise and speaker variability. The ablation study validates the critical role of combining these components, as the hybrid approach significantly enhances accuracy and generalizability across datasets, corroborating the design considerations outlined in table III.

E. Discussion

The model demonstrates strong overall performance in emotion recognition, particularly excelling in identifying anger (95%) and neutral (93%) states. The confusion matrix reveals minor challenges in distinguishing between certain emotion pairs, particularly fear-sad and disgust-sad. These confusions likely stem from acoustic similarities between these emotional expressions.

The training curves indicate effective learning progression, though the growing gap between training and validation metrics suggests potential overfitting. This observation points to opportunities for further optimization through additional regularization techniques or architectural refinements.

Despite these challenges, the model achieves a robust validation accuracy of 90.26%, demonstrating its effectiveness for practical speech emotion recognition applications. Future improvements could focus on:

- Enhanced feature engineering for better discrimination between similar emotions
- Implementation of additional regularization techniques
- Exploration of data augmentation strategies for challenging emotion pairs

FUTURE WORK

The current model is observed to have a fairly high accuracy of 90.26%, while avoiding bias learning or

overfitting into the dataset. The confusion matrix in Figure 3 shows high degrees of accuracy for each of the emotions analyzed. However, there are many possibilities for improvement that could be incorporated into the training of the model, thereby improving its performance.

A. Dataset

The performance of the model could be further improved by enhancing the dataset. This can be achieved by increasing the number of labeled audio samples to ensure a more diverse representation of each emotion. A larger dataset would help in better generalization and might help the model avoid potential bias due to underrepresentation of specific emotions. In addition, a diverse dataset should include data from multiple demographics of people for generalization.

B. Mel-frequency Cepstral Coefficients (MFCC)

MFCCs are features that represent the short-term power spectrum of sound, capturing perceptual characteristics of speech, such as pitch and tone. Exploring enhanced versions of MFCC could improve its performance in emotion recognition. For example, higher-order MFCC features, or combining MFCC with delta and delta-delta features, would capture the dynamic aspects of speech, such as changes in tone or intensity over time. Delta features track these changes, while delta-delta features capture second-order derivatives, which could be particularly useful for detecting subtle emotional shifts in speech.

C. Chroma Features

Chroma features capture the harmonic and melodic aspects of speech or music, representing pitch content that may reflect emotional states. These features can detect variations in tone and intensity, which are characteristic of emotions. For instance, higher-pitched speech often correlates with happiness or excitement, while lower-pitched speech may indicate sadness or anger. By incorporating chroma features alongside MFCC, the model can leverage both spectral and tonal information, enhancing its ability to recognize a broader range of emotional cues in speech. Furthermore, experimenting with different chroma features, such as chroma vector, chroma energy, and chroma difference, could reveal emotion-specific tonal patterns that may be missed by standard MFCC.

D. Model Architecture

Enhancing the model architecture is another critical avenue for improving performance. Future work could explore transformer-based models, which have demonstrated impressive results in various sequential tasks, potentially improving both accuracy and robustness. Transformers excel at capturing long-range dependencies in sequences, making them especially useful for modeling emotions that evolve over extended periods of time in speech. Moreover, fine-tuning pre-trained models and leveraging transfer learning could allow the model to benefit from large-scale pre-trained knowledge, leading to enhanced performance and faster convergence.

E. Cross-Domain Generalization

Evaluating the model's ability to generalize across different domains is crucial for robustness. Future work could train and test the model on diverse languages, cultures, and types of speech, such as spontaneous conversations versus scripted dialogues. This will ensure broader applicability, including multilingual environments. This would potentially help the model to adjust to new speakers, contexts, or emotional states not seen during training. Additionally, testing the model with varying background noise, microphone qualities, and emotional intensities will improve its real-world performance for applications like virtual assistants or customer service.

VI. Conclusion

In this study, we proposed a hybrid CNN-RNN architecture for Speech Emotion Recognition (SER) to address the challenges of capturing both spatial and temporal features inherent in emotional speech signals. By integrating convolutional layers for spatial feature extraction and recurrent layers for temporal sequence modeling, our approach demonstrated superior performance over traditional CNN-only and LSTM-only models, achieving a test accuracy of 90.26% on the combined dataset.

The results validate the efficacy of our architecture in overcoming the limitations of standalone methods, particularly in handling the dynamic nature of speech and noise resilience. Moreover, the comprehensive evaluation across multiple datasets—RAVDESS, TESS, and CREMA-D—highlights the model's robustness and generalizability. Future work can extend this research by exploring transformer-based models, advanced augmentation strategies, and real-time deployment to enhance SER systems further. This hybrid framework contributes a significant step towards more intuitive and effective human-computer interactions.

Acknowledgment

1) ChatGPT: Assisted in structuring the LaTeX document and organizing the content effectively.
2) Claude.AI: Helped in refining sentences and merging information cohesively from multiple sources.
3) Google Scholar: Used for sourcing relevant research papers related to Speech Emotion Recognition (SER).

REFERENCES

[1] S. G. Koolagudi, and K. S. Rao, "Emotion recognition from speech: a review," *International Journal of Speech Technology*, vol. 15, pp. 99–117, 2012. https://doi.org/10.1007/s10772-011-9125-1.

[2] B. Basharirad, and M. Moradhaseli, "Speech emotion recognition methods: A literature review," *AIP Conference Proceedings*, vol. 1891, no. 1, p. 020105, 2017. https://doi.org/10.1063/1.5005438.

[3] T. M. Wani, T. S. Gunawan, S. A. A. Qadri, M. Kartiwi, and E. Ambikairajah, "A Comprehensive Review of Speech Emotion Recognition Systems," *IEEE Access*, vol. 9, pp. 47795-47814, 2021. doi: 10.1109/ACCESS.2021.3068045.

[4] S. K. Pandey, H. S. Shekhawat, and S. R. M. Prasanna, "Deep Learning Techniques for Speech Emotion Recognition: A Review," in *2019 29th International Conference Radioelektronika (RADIOELEKTRONIKA)*, Pardubice, Czech Republic, 2019, pp. 1-6. doi: 10.1109/RADIOELEK.2019.8733432.

[5] M. Kalamani, S. Valarmathy, C. Poonkuzhali, and C. J. N., "Feature selection algorithms for automatic speech recognition," in *2014 International Conference on Computer Communication and Informatics*, Coimbatore, India, 2014, pp. 1-7. doi: 10.1109/ICCCI.2014.6921797.

[6] J. Nouza, "Feature selection methods for hidden Markov model-based speech recognition," in *Proceedings of 13th International Conference on Pattern Recognition*, Vienna, Austria, 1996, pp. 186-190 vol.2. doi: 10.1109/ICPR.1996.546749.

[7] C. Koniaris, M. Kuropatwinski, and W. B. Kleijn, "Auditory-model based robust feature selection for speech recognition," *Journal of the Acoustical Society of America*, vol. 127, no. 2, pp. EL73–EL79, Feb. 2010. https://doi.org/10.1121/1.3284545.

[8] N. Esfandian, F. Razzazi, and A. Behrad, "A clustering based feature selection method in spectro-temporal domain for speech recognition," *Engineering Applications of Artificial Intelligence*, vol. 25, no. 6, pp. 1194-1202, 2012. https://doi.org/10.1016/j.engappai.2012.04.004.

[9] K. J. Piczak, "Environmental sound classification with convolutional neural networks," in *2015 IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*, Boston, USA, Sept. 17–20, 2015. doi: https://doi.org/10.1109/MLSP.2015.7324337.

[10] T. N. Sainath, A.-r. Mohamed, B. Kingsbury, and B. Ramabhadran, "Deep convolutional neural networks for LVCSR," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Vancouver, Canada, May 2013, pp. 8614-8618. doi:10.1109/ICASSP.2013.6639347.

[11] K.-Y. Huang, C.-H. Wu, Q.-B. Hong, M.-H. Su, and Y.-R. Zeng, "Speech emotion recognition using convolutional neural network with audio word-based embedding," in *2020 IEEE International Conference on Multimedia and Expo (ICME)*, London, UK, July 2020, pp. 1-6. doi:10.1109/ICME46284.2020.9102814.

[12] S. Mirsamadi, E. Barsoum, and C. Zhang, "Automatic speech emotion recognition using recurrent neural networks with local attention," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, New Orleans, LA, USA, March 2017, pp. 2227-2231. doi:10.1109/ICASSP.2017.7952552.

[13] D. Li, J. Liu, Z. Yang, L. Sun, and Z. Wang, "Speech emotion recognition using recurrent neural networks with directional self-attention," *Expert Systems with Applications*, vol. 175, p. 114812, 2021. doi:10.1016/j.eswa.2021.114812.

[14] K. Choi, G. Fazekas, M. Sandler, and K. Cho, "Convolutional recurrent neural networks for music classification," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, New Orleans, LA, USA, March 2017, pp. 2392-2396. doi:10.1109/ICASSP.2017.7952906.

[15] 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. 1502-1506, Oct. 2018. doi:10.1109/LSP.2018.2869639.

[16] Bertero, D., Siddique, F., Wu, C.-S., Wan, Y., Chan, R., & Fung, P. (2016). Real-Time Speech Emotion and Sentiment Recognition for Interactive Dialogue Systems. 1042-1047. https://doi.org/10.18653/v1/D16-1110.

[17] R. A. Khalil, E. Jones, M. I. Babar, T. Jan, M. H. Zafar, and T. Alhussain, "Speech Emotion Recognition Using Deep Learning Techniques: A Review," in IEEE Access, vol. 7, pp. 117327-117345, 2019, doi: 10.1109/ACCESS.2019.2936124.

[18] S. Yoon, S. Byun, and K. Jung, "Multimodal Speech Emotion Recognition Using Audio and Text," 2018 IEEE Spoken Language Technology Workshop (SLT), Athens, Greece, 2018, pp. 112-118, doi: 10.1109/SLT.2018.8639583.

[19] Zhang, S., Yang, Y., Chen, C., Liu, R., Tao, X., Guo, W., Xu, Y., & Zhao, X. (2023). Multimodal emotion recognition based on audio and text by using hybrid attention networks. Biomedical Signal Processing and Control, 85, 105052. https://doi.org/10.1016/j.bspc.2023.105052.

[20] Mischler, G., Keshishian, M., Bickel, S., Mehta, A. D., & Mesgarani, N. (2023). Deep neural networks effectively model neural adaptation to changing background noise and suggest nonlinear noise filtering methods in auditory cortex. NeuroImage, 266, 119819. https://doi.org/10.1016/j.neuroimage.2022.119819.

[21] Eriş, F. G., & Akbal, E. (2024). Enhancing speech emotion recognition through deep learning and handcrafted feature fusion. Applied Acoustics, 222, 110070. https://doi.org/10.1016/j.apacoust.2024.110070.