Enhanced Deep Learning Techniques for Real-Time Speech Emotion Recognition in Multilingual Contexts

Donia Y. Badawood
Data Science Department, College of Computing, Umm Al-Qura University, Makkah, Saudi Arabia
dybadawood@uqu.edu.sa

Fahd M. Aldosari
Computer and Network Engineering Department, College of Computing, Umm Al-Qura University, Makkah, Saudi Arabia (corresponding author)
fmdosari@uqu.edu.sa

Received: 11 October 2024 | Revised: 26 October 2024 and 30 October 2024 | Accepted: 3 November 2024

ABSTRACT

Emotion recognition from speech is crucial for advancing human-computer interactions, enabling more natural and empathetic communication. This study proposes a novel Speech Emotion Recognition (SER) framework that integrates Convolutional Neural Networks (CNNs) and transformer-based architectures to capture local and contextual speech features. The model demonstrates strong classification performance, particularly for prominent emotions such as anger, sadness, and happiness. However, challenges persist in detecting less frequent emotions, such as surprise and calm, highlighting areas for improvement. The limitations of current datasets, such as limited linguistic diversity, are discussed. The findings underscore the model's robustness and identify avenues for future enhancement, such as incorporating more diverse datasets and employing techniques such as transfer learning. Future work will explore multimodal approaches and real-time implementation on edge devices to improve the system's adaptability in real-world scenarios.

Keywords-CNN; deep learning; speech emotion recognition; multilingual; real time

I. INTRODUCTION

Emotions play a vital role in communicating between humans, which profoundly affects understanding and the dynamics of interpersonal interactions. Speech Emotion Recognition (SER) has attracted considerable attention in recent years due to its promising applications in diverse areas, including customer service automation, mental health monitoring, human-computer interaction, and multimedia content analysis [1]. The implementation of SER has the potential to significantly improve user experience and satisfaction by allowing deeper understanding and response to human emotions, thus fostering more authentic and sensitive interactions [2]. The development of accurate and efficient SER systems is fraught with various challenges despite their potential advantages. Traditional approaches often fail to capture the intricate and subtle nature of emotional expressions in speech, relying mainly on manually crafted features and fundamental machine-learning techniques. The domain has experienced a notable evolution due to deep learning, which provides powerful techniques to automatically identify and extract pertinent features from unprocessed speech data [3-4]. Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), especially Long Short-Term Memory (LSTM) networks, have shown a lot of promise in finding patterns in voice data that are related to space and time. Current SER systems require enhancements to operate effectively in multilingual contexts, as they are designed mainly for monolingual applications. The obstacles encountered by SER systems striving for worldwide relevance are considerable, as variations in language and culture can profoundly affect the articulation and understanding of emotions [5, 6].

The proposed method incorporates sophisticated data augmentation strategies with a multilingual emotional lexicon to improve generalization and resilience across various languages and cultural settings. The objective is to enhance the model's ability to adjust to multiple languages and emotional subtleties by applying transfer learning and fine-tuning on various speech datasets. Extensive evaluations carried out on various benchmark datasets in multiple languages demonstrated the effectiveness of the proposed method, indicating enhancements in the accuracy of real-time information processing and emotion classification. This research introduces a method that combines sophisticated deep learning approaches with pragmatic aspects for multilingual applications. This represents a significant advance in the development of SER. The results provide significant implications for subsequent investigations and advancements in emotion-aware computing systems. The key contributions of this article are as follows:

- Proposes a deep learning framework for multilingual speech emotion recognition that can enhance human-computer interactions.
- Addresses linguistic and cultural challenges through multilingual data augmentation, making the model adaptable to diverse global environments.
- Utilizes advanced neural architectures, such as transformers and CNNs, for improved emotion detection accuracy in multilingual settings.
- It is designed for broad applications, including healthcare, virtual assistants, and customer service, promoting emotion-aware interactions in real-world contexts.

A. Speech Emotion Recognition (SER)

Automatic Speech Recognition (ASR) is a fundamental emotion indicator for fine-grained SER. In [7], the alignment lattice was proposed to enhance the ability to differentiate between emotional and indifferent frames. The blank sign was used within a transducer inference framework to develop the factorized Emotion Neural Transducer (ENT). Using the IEMOCAP dataset for utterance-level SER, the ENT models showed fewer word errors than the best methods in the field. Moreover, investigations on IEMOCAP and ZED, the latest speech emotion diarization dataset, have shown the effectiveness of fine-grained emotion modeling. In [8], five distinct categories of emotions, namely neutrality, happiness, anger, sadness, and excitement, were identified. After preprocessing the input voice data, the Dialogue Emotion Decoder (DED) was used to extract features and a CNN classifier was used to determine how the speakers felt about what they were saying. In [9], several studies on SER models utilizing CNNs were reviewed to identify and recommend the most effective methods for extracting emotions from speech data.

Most systems comprise three fundamental components, Data, Feature extraction, and Classification (DFC), to identify emotions conveyed through speech signals. This enhancement is expected to bolster the resilience of CNNs. In [10], the ASR output was integrated into the pipeline to facilitate joint training in SER. The integration of different ASR outputs and fusion methods was tested, particularly a hierarchical co-attention fusion approach, which significantly improved SER performance. This method achieved a weighted accuracy of 63.4% when analyzing the IEMOCAP corpus. In [11], a Speaker Recognition (SR) model was used, which had already been trained to perform Frontend Attribute Disentanglement (AD). The AD module comprised two stages, Attribution Reconstruction (AR) and Normalization (AN), which are critical for robust emotion discrimination. A dual space loss was proposed to enhance the separation of emotion-relevant and emotion-irrelevant spaces, improving the disentanglement process. In [12], the effectiveness of various combination methods using Multi-Task Learning (MTL) was examined, focusing on the significance of the style attribute. A selective multi-task learning approach was proposed and applied across all emotion categories except the neutral, demonstrating its effectiveness on the IEMOCAP database and a call center dataset. In [13], a method was presented to eliminate traditional feature extraction by directly processing the speech signal. This approach combined a spiking neural network (LSM) with the source-filter model of speech production. After processing, the neural reservoirs compressed the output, which was then classified for emotion recognition.

In [14], an innovative disentanglement network was introduced to separate the acoustic and emotional components. This method captured more nuanced and distinctive emotional signals, enhancing emotion detection accuracy by integrating identity-aware and disentanglement modules. In [15], a voice and speech emotion recognition system was proposed for emergency parking instructions, employing a Support Vector Machine (SVM) to classify emotions after extracting the feature vector from the voice signal.

B. Deep Learning Techniques

In [16], a study on classification tasks for ILSVRC 2015 was presented. Models with 100 and 1000 layers were examined using the CIFAR-10 dataset. This study demonstrated a performance enhancement of 28% on the COCO object identification dataset through the application of exceptionally deep networks. This approach resulted in first-place achievements in multiple categories of ILSVRC and COCO 2015, grounded in deep residual networks. In [17], the use of the Structural Similarity Index (SSIM) was proposed, showing how accurate it was by comparing it to other popular methods using subjective evaluations on a set of JPEG and JPEG2000-compressed images.

In [18], methods for handwriting character recognition were investigated, particularly highlighting the superiority of CNNs. This study examined the most important parts of document recognition systems and developed Graph Transformer Networks (GTNs) to improve multimodule systems by training them worldwide. In [19], the advantages of the proposed architecture in efficiently utilizing computational resources were highlighted. Multiscale processing and the Hebbian principle affected them, as they increased the network's width and depth while maintaining a constant computational cost. In [20], end-to-end integration within a single network architecture enhanced detection performance, particularly with YOLO. This efficiency was demonstrated by processing images in real-time at an impressive rate, showcasing its superior mean Average Precision (mAP) compared to other real-time detectors. In [21] the Region Proposal Network (RPN) was introduced to facilitate region proposals at minimal costs. In [22], fully convolutional networks were used to learn from inputs of different sizes, especially for predictions that are dense in space. Classification networks such as AlexNet and GoogLeNet were fine-tuned for segmentation tasks, enhancing segmentation precision through a skip architecture. In [23], the impact of computational steps on pedestrian detection performance was demonstrated by a method that outperformed other techniques on the MIT pedestrian database.

In [24], the Swin transformer was presented, a hierarchical vision transformer designed for visual tasks. A shifted windowing strategy was introduced, which improved efficiency by confining self-attention calculations to local windows while maintaining global connections. In [25], NSGA-II was introduced, which is a Multi-Objective Evolutionary Algorithm (MOEA) that addresses challenges in previous methods. A fast, non-dominated sorting algorithm was developed and a population diversity and fitness selection operator was introduced.

Several previous investigations have sought to address the challenges, with some employing transformer-based models to enhance contextual comprehension. However, the majority of current research emphasizes monolingual datasets or fails to incorporate a cohesive integration of local and global speech features, restricting their applicability in practical multilingual contexts. Furthermore, a significant number of SER models lack optimization for real-time applications, an essential requirement for interactive systems.

The proposed method effectively addresses existing gaps by integrating CNNs for local feature extraction alongside a transformer-based encoder, which is adept at capturing global contextual information. This hybrid architecture facilitates enhanced emotion classification, particularly in multilingual and real-time contexts. The implementation of multilingual data augmentation significantly improves model generalization across various languages, a characteristic that is frequently absent in current methods. Furthermore, this study examines the practical challenges associated with real-time deployment, highlighting the importance of optimizing the model for edge devices and real-world applications.

II. METHODOLOGY

This methodology outlines a systematic approach to employing deep learning techniques tailored for multilingual contexts in real-time SER. Figure 1 shows the process involved in the real-time SER analysis using a deep learning framework.

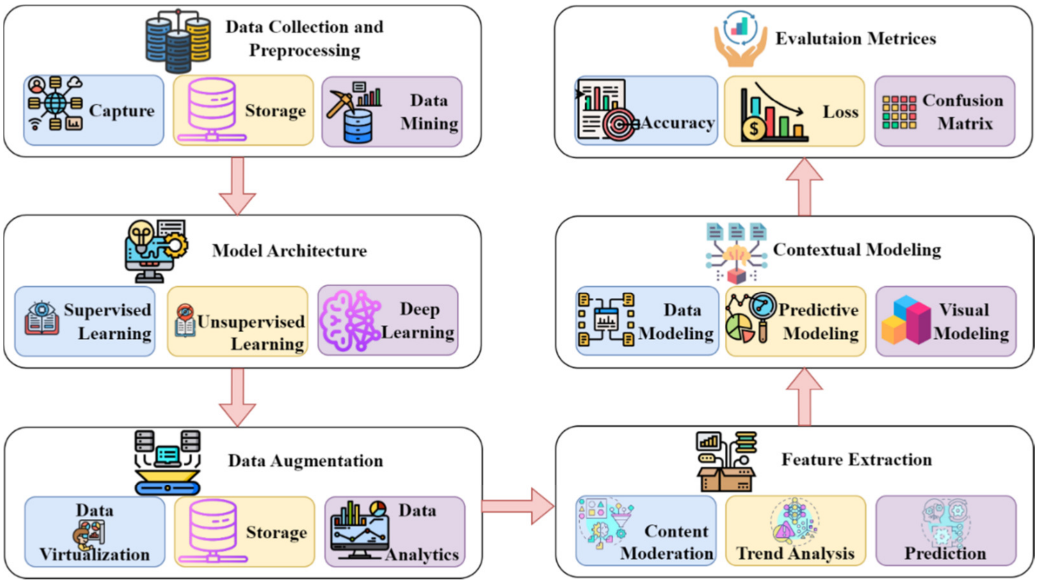


Fig. 1. Process of real-time multilingual SER.

A. Data Collection and Preprocessing

Utilizing diverse speech datasets that encompass various languages, emotional states, and speaker demographics is crucial to developing an effective SER system. This diversity ensures robust model generalization across various situations and user demographics. This research utilizes several significant datasets. The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) [26, 27], comprises a collection of recordings that encapsulate eight distinct emotions in English, including anger, disgust, fear, happiness, neutrality, sadness, surprise, and calmness, represented through both speech and musical formats. The CREMA-D dataset [27] contains vocal emotional expressions under six labels: happy, sad, anger, fear, disgust, and neutral.

The Surrey Audio-Visual Expressed Emotion (SAVEE) dataset [28] consists of a significant compilation of expressive speech and facial expressions, meticulously curated to support SER research. This dataset consists of audio recordings featuring actors conveying seven unique emotions: anger, disgust, fear, happiness, sorrow, surprise, and neutral, and enables a thorough examination of emotion analysis through the integration of audio and video recordings. The variety of expressions used to communicate different emotions offers significant data that can be used for the training and evaluation of emotion detection systems. The SAVEE dataset serves as a significant resource for the advancement of emotion detection models, characterized by its balanced representation of various emotions and speaker demographics. The Toronto Emotional Speech Set (TESS) [29] represents a comprehensive dataset that holds significant potential for advancing research in the

SER domain. The compilation consists of audio recordings of female performers delivering emotionally impactful speeches in English. The dataset encompasses a range of emotional states, including anger, disgust, fear, happiness, sadness, surprise, and neutrality.

Every emotion is defined by a range of utterances, ensuring a thorough array of emotional expressions and speech patterns. TESS is a valuable asset for training and evaluating SER systems, due to its superior quality recordings and well-defined emotional labels. Advancing models to accurately detect and interpret emotions conveyed through speech is paramount. It is essential to implement specific preprocessing procedures to ensure that the input data are pristine, reliable, and appropriate for model training. Noise reduction algorithms were employed to improve the clarity of speech signals and reduce the impact of background noise. Segmentation entails partitioning speech recordings into smaller, more manageable units, guided by natural boundaries such as utterances or emotive words, enabling a more focused analysis. Normalization is used to standardize audio signals, achieve consistency in loudness and pitch levels, and reduce the variability introduced by recording conditions. Feature extraction involves obtaining pertinent acoustic characteristics, such as pitch, energy, and Mel-Frequency Cepstral Coefficients (MFCCs). The features function as inputs for the deep learning models and play a crucial role in capturing the fundamental characteristics of the spoken signal. These preprocessing procedures aimed at enhancing the development of a high-performing SER system by optimizing data for efficient training and evaluation.

B. Model Architecture and Data Augmentation

The proposed approach integrates a transformer-based architecture with CNNs to effectively capture both local and contextual features inherent in voice signals, seeking to improve the strength and precision of SER. To enhance the reliability and utility of the model, a variety of data augmentation techniques were used, which simulate diverse recording environments and speaker variations. Among these methods, noise injection stands out as a key approach, where various background noise levels are integrated into speech recordings to simulate various environmental conditions, from bustling streets to tranquil chambers. This method allows the model to withstand real-world conditions where background noise may hinder speech clarity and recognition precision. Pitch shifting represents a distinctive modification implemented to address variations in speaker characteristics, including age, gender, and vocal tone, by altering the pitch within speech recordings. The implementation of pitch shifting enhances the model's ability to adapt to various voice profiles, improving its accuracy in emotion recognition across a wide range of speakers.

Speed variation involves modifying the speed of speech recordings, facilitating the simulation of various speaking rates and, consequently, producing a more diverse training dataset. This approach improves the model's ability to identify emotions in speech delivered at different tempos, enhancing its applicability across various speech patterns and speeds. Incorporating these techniques into a dataset results in an expanded training set, enhancing the model's applicability across a broader spectrum of real-world scenarios and speaker variations. A CNN-based model examines the raw voice signal during the feature extraction phase. This module integrates several convolutional layers to extract prosody and spectral patterns, which are essential local auditory variables to evaluate emotional states. CNNs exhibit a significant capacity to recognize intricate patterns in the speech stream, encompassing pitch, tone, and intensity variations. These characteristics offer a detailed depiction of the speech signal, establishing a basis for further examination in the following phases.

The system employs a transformer-based encoder to collect global contextual information and interpret the emotional subtleties present in the speech. The transformer encoder holds the features derived from the CNN to examine the relationships between various speed training, employing an attention mechanism to enhance contextual understanding. This method enables the model to discern the relationships between different speech signal components, thus enhancing comprehension of the emotional context. Implementing a transformer-based encoder significantly improves the model's capacity to articulate and distinguish between diverse linguistic and emotional contexts by emphasizing these contextual dependencies.

C. Training, Evaluation, and Performance Metrics

The model training phase utilizes a systematic method to ensure generalization and optimize performance. The method for emotion classification employs cross-entropy loss as the main objective function. This approach accurately quantifies the difference between the observed and anticipated emotional classifications. Additional regularization strategies, including dropout and weight decay, are integrated during the training phase to improve model robustness and reduce the likelihood of overfitting. In the dataset allocation process, 80% is allocated for the training phase, while the remaining 20% is reserved for validation. This division facilitates practical hyperparameter tuning and model assessment during the training phase. Random and grid search techniques were employed to identify the optimal set of hyperparameters, ensuring that the model performs effectively across various conditions.

D. Dataset Limitations

The datasets utilized, RAVDESS, CREMA-D, SAVEE, and TESS, are well-established resources within the SER domain. However, each of them exhibits specific limitations that influence the generalizability of models developed using them. The RAVDESS dataset exhibits a variety of emotional expressions but is predominantly composed of English-language recordings, which constrains the model's capacity to generalize to other languages. In a similar vein, the CREMA-D dataset presents a diverse array of emotions and speakers but, again, it is limited to English, which constrains the model's utility in multilingual scenarios. The SAVEE dataset offers significant resources for the study of emotion detection, integrating both audio and visual elements. However, similar to the earlier datasets, it exhibits a deficiency in linguistic diversity, as all recordings are exclusively in English. Furthermore, the restricted number of speakers in certain datasets may lead to bias during both the training and testing phases, ultimately diminishing the robustness of the models when applied to more varied real-world populations. The TESS dataset provides high-quality emotional speech recordings, but its exclusive use of female speakers presents limitations that may hinder the model's capacity to generalize across various genders. The focus of the dataset on a limited spectrum of speakers could potentially constrain the model's efficacy when it encounters a broader array of voices in practical scenarios.

III. RESULTS AND DISCUSSION

Figure 2 illustrates the dataset's distribution of emotions, providing a comprehensive understanding of the model's performance. The figure displays the frequency of each emotional category that the proposed SER model predicted, presenting a clear and concise representation of the frequency of various emotions identified by the model, including anger, happiness, sorrow, fear, surprise, disgust, and neutrality.

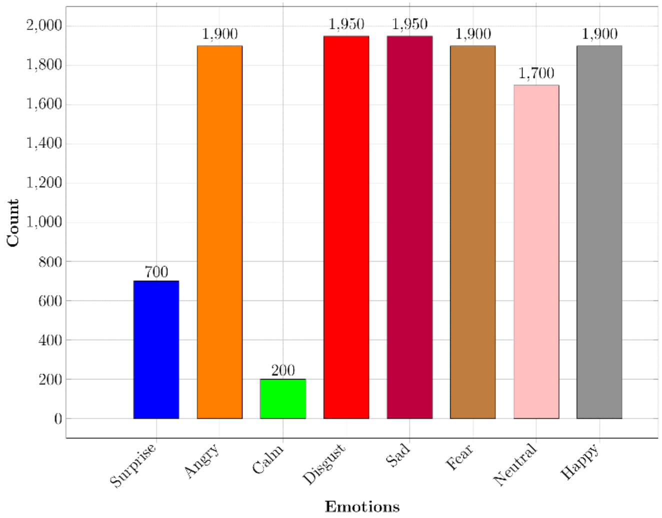


Fig. 2. Quantification of affective states.

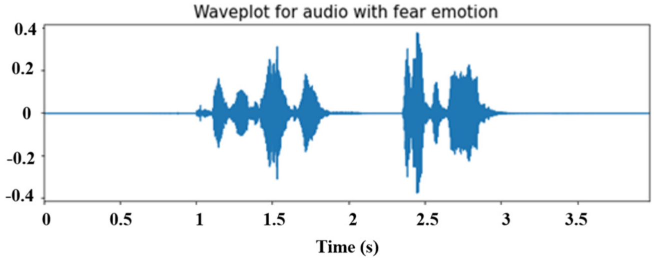


Fig. 3. Waveform representation for audio exhibiting the emotion of fear.

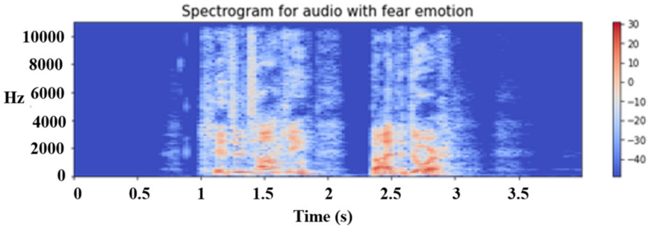


Fig. 4. Spectrogram analysis of audio exhibiting the fear emotion.

Figure 3 shows a visual representation of the amplitude of a signal communicated over time corresponding to audio samples associated with fear emotion. This graph depicts the variations in the audio waveform associated with fearful events, highlighting the unique rhythm, pitch, and intensity fluctuations related to this emotional state. To make the model better at classifying emotions, it is important to look at the unique sound features of terrified speech, such as increased pitch variation and dynamic amplitude fluctuations, using wave plot analysis. This visual assistance enhances pattern recognition and assesses the effectiveness of the proposed SER model in capturing and distinguishing emotional cues related to fear. Figure 4 shows the frequency spectrum of the signal emitted over time for audio samples that convey fear emotion. This graphic analyzes the audio's frequency content alterations, emphasizing trends such as increased frequency modulation sharpness and heightened energy within particular frequency ranges. The spectrogram illustrates the auditory features of terrified speech, characterized by increased intensity variations and more pronounced and erratic spectral shifts. Examining these patterns can reveal more about how fear changes the way people talk, which will help the model better recognize and label fear-related emotions in a wide range of audio samples.

The wave plot in Figure 5 illustrates the temporal variations in the speech signal amplitude for audio samples classified as melancholy. Compared to other emotional expressions, the observed amplitude patterns typically exhibit a smoother and less varied profile, indicative of the diminished intensity and subdued characteristics associated with sad speech. The bars illustrate the model's capacity to recognize and differentiate among various emotional states, each representing the total number of instances categorized into the specific emotion.

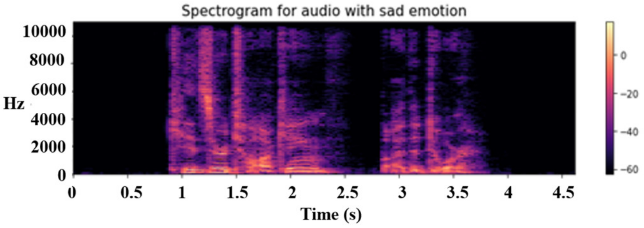


Fig. 5. Wave plot corresponding to audio characterized by a sad emotional expression.

This investigation highlights the identification of both the advantages and disadvantages associated with the model. For instance, a higher frequency of occurrence of any particular emotion within the model's output might imply an improved capacity to recognize particular emotional states. Reduced frequencies for alternative emotions can suggest specific domains where the model requires additional training or data enhancement. The distribution of emotions allows evaluation of the model's equity and balanced predictions across different emotional categories. These data are essential for evaluating the model's performance in real-world scenarios where accurate and unbiased emotion recognition is necessary. Gradual alterations in the waveform and the reduction in intensity indicate a somber emotional tone. To ensure the precise recognition and classification of sad speech within the SER system, it is essential to define its acoustic characteristics. This can be accomplished by a detailed analysis of the wave plot.

The spectrogram of audio recordings in Figure 6 illustrating depression reveals the frequency content and its temporal progression. Compared to other emotional states, this representation typically exhibits a more subdued and stable frequency spectrum characterized by reduced frequency modulation. The subdued and more uniform characteristics of melancholy speech are evident in the reduced intensity and nuanced spectral variations. Studying these spectral patterns provides details on how depression changes the way people sound when they speak. This makes it easier to find and label people who are depressed in a variety of audio samples.

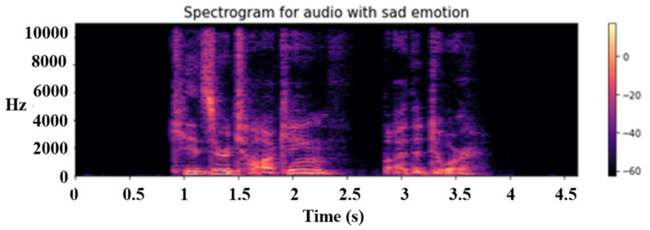


Fig. 6. Spectrogram corresponding to audio characterized by a sad emotional expression.

The audio samples associated with happiness exhibit distinct waveform characteristics and notable dynamic amplitude fluctuations within their wave plots, as shown in Figure 7. The energy and enthusiasm in speech indicate happiness, often observable in the associated plot through notable fluctuations and increased peaks.

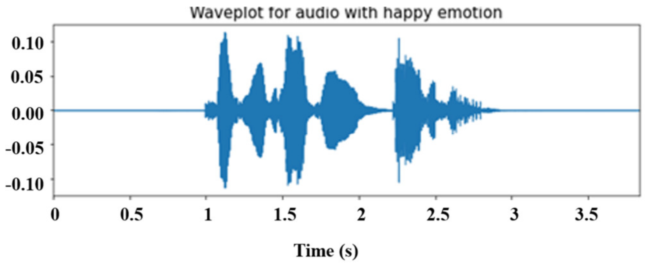


Fig. 7. Waveform representation for audio exhibiting a happy emotional state.

Distinct amplitude and rhythmic patterns are associated with happiness's animated and expressive qualities. Examining these amplitude shifts, valuable insights can be derived concerning the impact of positive emotional states on speech signals, which can improve the SER system's capacity to accurately identify emotions. The frequency patterns identified in the spectrogram for the audio samples indicative of happiness demonstrate a vibrant and dynamic quality, as shown in Figure 8.

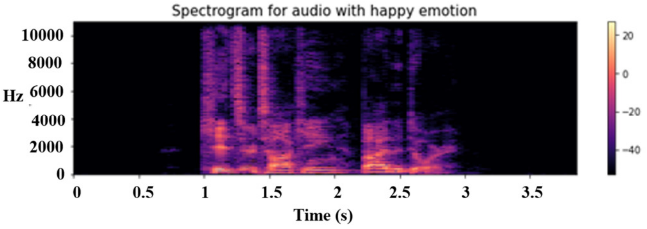
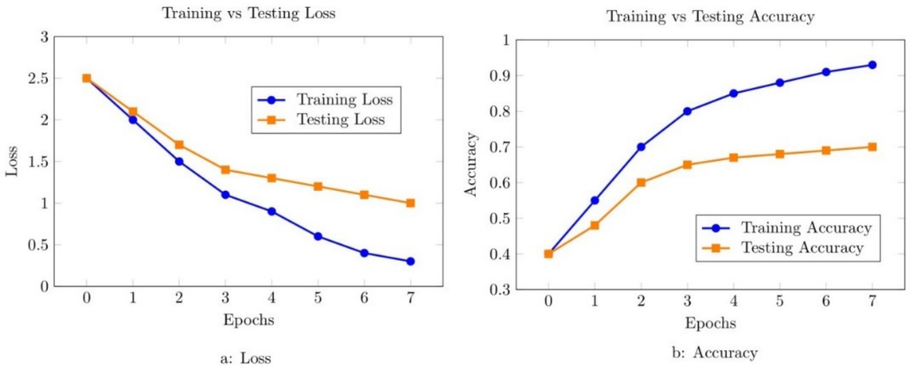


Fig. 8. Analysis of spectrogram representations for audio exhibiting happy emotional states.

The features of joyful speech generally encompass notable frequency modulation and vibrant vocal expressions. Joyful emotions display unique intonation patterns and rhythmic variations in frequency bands, frequently displaying brighter and more diverse spectral characteristics. By examining these patterns, a more profound understanding of the connection between happiness and vocal expressiveness can be achieved, thus improving the ability of speech analysis to identify emotional states. Figure 9 presents the accuracy and loss curves for training and testing the proposed model throughout the epochs. The training process is illustrated through a visual representation of the model's performance. Accuracy curves offer valuable insights into the model's predictive performance about correct labels. The accuracy observed during testing reflects the model's capacity to generalize to data it has not encountered before, while training accuracy generally demonstrates enhancement as the model gains experience. The loss curves depict the error rates recorded throughout the training and testing phases, where a decrease in loss signifies an enhancement in model performance. Evaluating these curves is crucial for comprehending the model's convergence, identifying possible overfitting, and assessing the overall efficacy of the learning process.



a: Loss

b: Accuracy

Fig. 9. Evaluation Metrics: Training and testing accuracy and loss.

The model's performance across different emotional categories was evaluated through a comprehensive analysis utilizing the confusion matrix illustrated in Figure 10. The matrix presents a detailed distribution of emotions, documenting 999 occurrences of Sadness, 999 instances of Calm, 122 instances of Anger, 701 instances of Disgust, 738 instances of Fear, 898 instances of Happy, 724 instances of Neutral, 999 instances of Sadness, and 392 instances of Surprise. The matrix comprises cells denoting the predicted instances corresponding to each emotion label. The model successfully predicted 967 instances of rage, whereas it only achieved 122 predictions of calm feelings. The analysis of the 999 and 898 cases shows the model's notable effectiveness in distinguishing between happiness and melancholy. The confusion matrix offers essential insights regarding the model's accuracy and potential biases within emotional categories, highlighting its strengths and areas that require enhancement in its capacity to distinguish between different emotions.

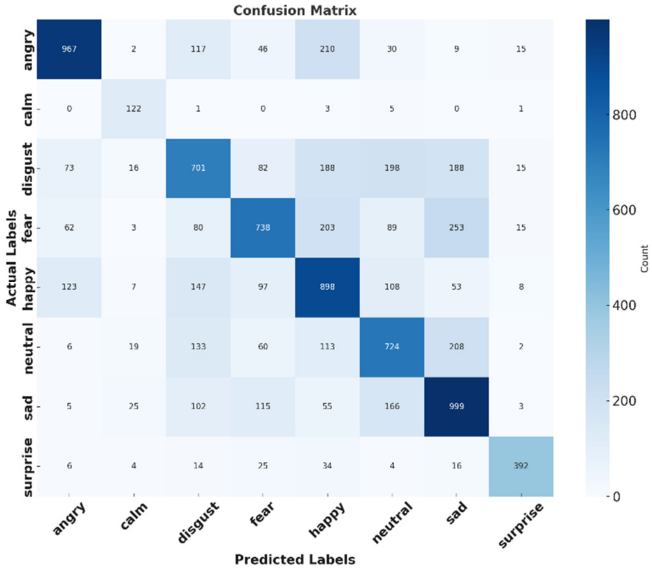


Fig. 10. Confusion matrix for analysis of various emotional states.

A. *Challenges and Limitations*

Due to dataset imbalance, the proposed model encountered challenges in identifying less prevalent emotions, such as surprise and calm. This imbalance skews the model's performance toward more frequently represented emotions, including anger and happiness. This constrains its ability to generalize across less common emotional expressions. Furthermore, the dependence on primarily English-language datasets limits the model's performance in multilingual environments, where it faces challenges in generalizing across various languages, accents, and dialects, even with attempts to improve diversity through data augmentation methods.

The deployment of models in real-time on-edge devices introduces significant challenges, necessitating additional optimization to address resource limitations such as computational power and latency encountered in practical applications. Furthermore, focusing solely on audio data constrains the model's efficacy, as incorporating multimodal input, such as visual or textual information, can significantly improve emotion recognition. Mitigating these limitations in subsequent research will enhance the system's adaptability and overall efficacy across various practical applications.

B. *Future Works*

Critical areas for enhancement involve integrating few-shot learning and transfer learning to improve the model's ability to identify rare emotions with limited data. Few-shot learning can enhance generalization from a limited number of examples of rare emotions. In contrast, transfer learning could utilize pre-trained models from various domains, boosting performance in multilingual and low-resource environments. Investigating multimodal approaches, including integrating visual or textual information with audio, can enhance emotion detection by effectively capturing intricate emotional cues that may not be easily discernible through audio alone. Implementing real-time systems on edge devices presents significant challenges, primarily due to inherent resource limitations. It is essential to implement optimization techniques, such as model compression, pruning, and quantization, to address this issue, as these methods aim to decrease latency and improve computational efficiency while maintaining accuracy.

IV. CONCLUSION

The proposed SER system incorporates transformer-based architecture with a CNN. The findings suggest that the model exhibits significant effectiveness in emotion classification, successfully combining both the local and contextual aspects of speech. The confusion matrix indicated a high proficiency in emotion recognition, particularly for anger, sadness, and happiness. However, the model exhibited difficulties in recognizing less common emotions, such as surprise and calmness, highlighting particular areas that require additional improvement. The model's convergence and generalization capacity were evaluated by analyzing the training and testing accuracy and loss curves. Wave plots and spectrograms yielded significant insight into the auditory characteristics associated with various emotions. To improve the model's resilience in various contexts, future studies should expand the dataset to include a wider array of languages and emotional expressions.

Implementing sophisticated methods such as few-shot learning or transfer learning can significantly improve the model's capacity to identify rare emotional states with limited data input. Investigating multimodal approaches that combine auditory and textual signals can improve emotion recognition accuracy and offer a more profound insight into emotional states. Implementing real-time solutions and optimization for diverse edge devices can ensure that applications function effectively within interactive systems and real-world environments.

REFERENCES

[1] 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, https://doi.org/10.1109/ACCESS.2021.3068045.

[2] M. Spezialetti, G. Placidi, and S. Rossi, "Emotion Recognition for Human-Robot Interaction: Recent Advances and Future Perspectives,"

Frontiers in Robotics and AI, vol. 7, Dec. 2020, https://doi.org/10.3389/frobt.2020.532279.

[3] D. Issa, M. Fatih Demirci, and A. Yazici, "Speech emotion recognition with deep convolutional neural networks," *Biomedical Signal Processing and Control*, vol. 59, May 2020, Art. no. 101894, https://doi.org/10.1016/j.bspc.2020.101894.

[4] R. Damaševičius, S. K. Jagatheesaperumal, R. N. V. P. S. Kandala, S. Hussain, R. Alizadehsani, and J. M. Gorriz, "Deep learning for personalized health monitoring and prediction: A review," *Computational Intelligence*, vol. 40, no. 3, 2024, Art. no. e12682, https://doi.org/10.1111/coin.12682.

[5] M. B. Akçay and K. Oğuz, "Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers," *Speech Communication*, vol. 116, pp. 56–76, Jan. 2020, https://doi.org/10.1016/j.specom.2019.12.001.

[6] E. K. Zadeh and M. Alaeifard, "Adaptive Virtual Assistant Interaction through Real-Time Speech Emotion Analysis Using Hybrid Deep Learning Models and Contextual Awareness," *International Journal of Advanced Human Computer Interaction*, vol. 1, no. 1, pp. 1–15, Jul. 2023.

[7] S. Shen, Y. Gao, F. Liu, H. Wang, and A. Zhou, "Emotion Neural Transducer for Fine-Grained Speech Emotion Recognition," in *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Seoul, Korea, Republic of, Apr. 2024, pp. 10111–10115, https://doi.org/10.1109/ICASSP48485.2024.10446974.

[8] S. N. Atkar, R. Agrawal, C. Dhule, N. C. Morris, P. Saraf, and K. Kalbande, "Speech Emotion Recognition using Dialogue Emotion Decoder and CNN Classifier," in *2023 2nd International Conference on Applied Artificial Intelligence and Computing (ICAAIC)*, Salem, India, May 2023, pp. 94–99, https://doi.org/10.1109/ICAAIC56838.2023.10141417.

[9] S. Malla, A. Alsadoon, and S. K. Bajaj, "A DFC taxonomy of Speech emotion recognition based on convolutional neural network from speech signal," in *2020 5th International Conference on Innovative Technologies in Intelligent Systems and Industrial Applications (CITISIA)*, Sydney, Australia, Nov. 2020, pp. 1–10, https://doi.org/10.1109/CITISIA50690.2020.9371841.

[10] Y. Li, P. Bell, and C. Lai, "Fusing ASR Outputs in Joint Training for Speech Emotion Recognition," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Singapore, May 2022, pp. 7362–7366, https://doi.org/10.1109/ICASSP43922.2022.9746289.

[11] Y. X. Xi, Y. Song, L. R. Dai, I. McLoughlin, and L. Liu, "Frontend Attributes Disentanglement for Speech Emotion Recognition," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Singapore, Singapore, May 2022, pp. 7712–7716, https://doi.org/10.1109/ICASSP43922.2022.9746691.

[12] H. Zhang, M. Mimura, T. Kawahara, and K. Ishizuka, "Selective Multi-Task Learning For Speech Emotion Recognition Using Corpora Of Different Styles," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Singapore, May 2022, pp. 7707–7711, https://doi.org/10.1109/ICASSP43922.2022.9747466.

[13] R. Lotfidereshgi and P. Gournay, "Biologically inspired speech emotion recognition," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, New Orleans, LA, USA, Mar. 2017, pp. 5135–5139, https://doi.org/10.1109/ICASSP.2017.7953135.

[14] Z. Yuan, C. L. Philip Chen, S. Li, and T. Zhang, "Disentanglement Network: Disentangle the Emotional Features from Acoustic Features for Speech Emotion Recognition," in *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Seoul, Republic of Korea, Apr. 2024, pp. 11686–11690, https://doi.org/10.1109/ICASSP48485.2024.10448044.

[15] T. Kexin, H. Yongming, Z. Guobao, and Z. Lin, "Research on Emergency Parking Instruction Recognition Based on Speech Recognition and Speech Emotion Recognition," in *2019 Chinese Automation Congress (CAC)*, Hangzhou, China, Nov. 2019, pp. 2933–2937, https://doi.org/10.1109/CAC48633.2019.8997077.

[16] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, Jun. 2016, pp. 770–778, https://doi.org/10.1109/CVPR.2016.90.

[17] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image Quality Assessment: From Error Visibility to Structural Similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, Apr. 2004, https://doi.org/10.1109/TIP.2003.819861.

[18] D. Parres and R. Paredes, "Fine-Tuning Vision Encoder–Decoder Transformers for Handwriting Text Recognition on Historical Documents," in *Document Analysis and Recognition - ICDAR 2023*, San José, CA, USA, 2023, pp. 253–268, https://doi.org/10.1007/978-3-031-41685-9_16.

[19] C. Szegedy *et al.*, "Going deeper with convolutions," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Boston, MA, USA, Jun. 2015, pp. 1–9, https://doi.org/10.1109/CVPR.2015.7298594.

[20] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You Only Look Once: Unified, Real-Time Object Detection," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, Jun. 2016, pp. 779–788, https://doi.org/10.1109/CVPR.2016.91.

[21] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017, https://doi.org/10.1109/TPAMI.2016.2577031.

[22] E. Shelhamer, J. Long, and T. Darrell, "Fully Convolutional Networks for Semantic Segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 4, pp. 640–651, Apr. 2017, https://doi.org/10.1109/TPAMI.2016.2572683.

[23] N. Dalal and B. Triggs, "Histograms of Oriented Gradients for Human Detection," in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, San Diego, CA, USA, 2005, vol. 1, pp. 886–893, https://doi.org/10.1109/CVPR.2005.177.

[24] Z. Liu *et al.*, "Swin Transformer: Hierarchical Vision Transformer using Shifted Windows," in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, Montreal, Canada, Oct. 2021, pp. 9992–10002, https://doi.org/10.1109/ICCV48922.2021.00986.

[25] K. Deb, A. Pratap, S. Agarwal, and T. Meyarivan, "A fast and elitist multiobjective genetic algorithm: NSGA-II," *IEEE Transactions on Evolutionary Computation*, vol. 6, no. 2, pp. 182–197, Apr. 2002, https://doi.org/10.1109/4235.996017.

[26] M. Sharafi, M. Yazdchi, R. Rasti, and F. Nasimi, "A novel spatio-temporal convolutional neural framework for multimodal emotion recognition," *Biomedical Signal Processing and Control*, vol. 78, Sep. 2022, Art. no. 103970, https://doi.org/10.1016/j.bspc.2022.103970.

[27] B. Mocanu, R. Tapu, and T. Zaharia, "Multimodal emotion recognition using cross modal audio-video fusion with attention and deep metric learning," *Image and Vision Computing*, vol. 133, May 2023, Art. no. 104676, https://doi.org/10.1016/j.imavis.2023.104676.

[28] L. Alhinti, S. Cunningham, and H. Christensen, "The Dysarthric Expressed Emotional Database (DEED): An audio-visual database in British English," *PLOS ONE*, vol. 18, no. 8, 2023, Art. no. e0287971, https://doi.org/10.1371/journal.pone.0287971.

[29] P. T. Krishnan, A. N. Joseph Raj, and V. Rajangam, "Emotion classification from speech signal based on empirical mode decomposition and non-linear features," *Complex & Intelligent Systems*, vol. 7, no. 4, pp. 1919–1934, Aug. 2021, https://doi.org/10.1007/s40747-021-00295-z.