IBIMA Publishing
Communications of International Proceedings
https://ibimapublishing.com/p-articles/45AI/2025/4519925/
Vol. 2025 (20), Article ID 4519925

Humanizing AI Chatbots: The Role of Speech Emotion Recognition with Deep Learning*

Bashar BARMADA, Madawa Gihan KANNANGARA, Guillermo RAMIREZ-PRADO, Soheil POUR and Masoud SHAKIBA

Unitec, Auckland, New Zealand

Correspondence should be addressed to: Bashar BARMADA, bbarmada@unitec.ac.nz

* Presented at the 45th IBIMA International Conference, 25-26 June 2025, Cordoba, Spain

Abstract

This research focuses on integrating Speech Emotion Recognition (SER) with AI chatbots to create a system that is more emotionally intelligent and responsive. Using advanced deep learning techniques such as Convolutional Neural Networks (CNNs), the study enhances the accuracy and robustness of SER models in detecting emotions from speech. The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) serves as the primary dataset, with data augmentation techniques such as noise injection, speed variation, and pitch shifting applied to improve model performance. Key features such as Mel-Frequency Cepstral Coefficients (MFCC), Mel Spectrogram, and Zero Crossing Rate (ZCR) are extracted to improve the analysis. The study uses regularization techniques, including Batch Normalization and L2 Regularization, to prevent overfitting. Eight emotion classes are evaluated namely, neutral, calm, happy, sad, angry, fear, disgust and surprise. Experimental results show significant improvements with the best test accuracy reaching 87.5%, outperforming previous studies. Visualized training history demonstrates the model's learning behavior and generalization capabilities. The findings highlight the potential of SER-enhanced chatbots in applications such as customer service and mental health support by enabling empathetic interactions.

Keywords: Speech emotion recognition, CNN deep learning, AI chatbot.

Introduction

Artificial intelligence (AI) is advancing in various domains, aiming to create more human-like interactions. One of the most remarkable fields in AI is the integration of Speech Emotion Recognition (SER) capabilities into AI chatbots. The technology aims to enhance the emotional intelligence of AI systems, making interactions with machines more engaging and natural.

The integration of SER into AI chatbots and other interactive systems has a significant potential, making human-computer interactions more personalized, empathetic, and efficient (Kapoor, 2022; Tariq, 2019). SER enables the chatbot to recognize and interpret human emotions from speech, improving applications in customer service, mental health support, and personal assistance by providing more empathic responses, demonstrating the role of SER in

Cite this Article as: Bashar BARMADA, Madawa Gihan KANNANGARA, Guillermo RAMIREZ-PRADO, Soheil POUR and Masoud SHAKIBA, Vol. 2025 (20) "Humanizing AI Chatbots: The Role of Speech Emotion Recognition with Deep Learning " Communications of International Proceedings, Vol. 2025 (20), Article ID 4519925, https://doi.org/10.5171/2025.4519925

improving human-technology interactions. In customer service, chatbots with SER capabilities can detect frustration or confusion in a user's tone, allowing for proactive, empathetic responses that improve customer satisfaction. Similarly, in mental health support, emotion-aware AI systems can identify signs of distress or anxiety, providing timely interventions or alerts to caregivers. Another promising application is in education, where SER-powered virtual tutors can adjust tone and content delivery based on the student's emotional state, creating a more engaging and adaptive learning experience (Abdelhamid, 2019). By recognizing and interpreting emotional cues from speech, the integration of SER with AI chatbots will lead to improved user satisfaction and trust.

This research explores the integration of SER in AI chatbots to enhance emotional intelligence. The study builds on previous work by Mustaqeem and Kwon (Mustaqeem, 2019), and Kapoor and Kumar (Kapoor, 2022), aiming to improve the accuracy and robustness of SER models. It uses advanced deep learning techniques, particularly Convolutional Neural Networks (CNNs), to enhance feature extraction and temporal analysis. Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) serves as the primary dataset, enabling a diverse range of emotional speech samples. The research highlights the impact of emotionally aware AI on user interactions, enabling chatbots to detect emotions such as happiness, sadness, anger and fear, and generate more human-like responses. The study contributes to the advancement of SER by developing more accurate and reliable models, paving the way for AI chatbots with enhanced emotional intelligence.

The Evolution of Speech Emotion Recognition

The world of Speech Emotion Recognition (SER) has made significant progress recently, thanks to the integration of deep learning methods and innovative ways to extract features. Convolutional Neural Network (CNN) has been used intensively to enhance audio signal processing and detect emotions, such as anger and stress, by combining traditional features with deep-learned features from speech spectrograms (Kapoor, 2022; Mustaqeem, 2019; Choudhary, 2022; Issa, 2020). CNN is also combined with other advanced algorithms, such as transformer-based models (Xu, 2021; Ullah, 2023) and Long Short-Term Memory (LSTM) networks (Abdelhamid, 2022), to achieve high precision of SER systems. Another way to improve the performance of SER is to combine the audio source with visual cues (Sultana, 2022; Dedeoglu, 2019), or with text sources (Singh, 2021) to provide additional information in training the CNN model.

Pre-processing and feature extraction methods play an important role in the building of SER systems. (Farooq, 2020) investigated how different feature selection algorithms impact CNN-based emotion recognition. They highlighted the importance of choosing appropriate features to optimize emotion recognition model performance. (Aggarwal, 2022) showed the importance of designing features using two-way feature extraction to improve the accuracy of emotion recognition systems. Temporal and spatial features of the audio could be a crucial component of the SER model (An, 2021). Dual Feature Extraction Encoders have also emerged as an approach to refining SER systems (Pulatov, 2023). The integration of deep learning with audio features such as Mel Frequency Cepstral Coefficients (MFCCs) has led to enhance the accuracy of emotions recognition (Hazra, 2022; Gupta, 2021).

Many applications have demonstrated the benefits of embedding AI in SER. (Tariq, 2019) proposed an approach to using deep learning for health care based on IoT to detect emotions from speech to determine the impact on patient well-being. (Naas, 2020) investigated real-time emotion recognition in sales environments to expand customer interactions and decision-making processes. AI in SER is also applied to capture emotions in multilingual audio to discover intricate emotional nuances expressed in various languages (Sultana, 2022; Bhattacharya, 2022). Natural Language Processing (NLP) applications can benefit from embedding AI in SER model to build conversational AI, as shown in (Johnson, 2021).

The studies above highlight the evolving landscape of speech emotion recognition from healthcare applications to innovative feature extraction methods and multimodal approaches. Various paths are explored to create more accurate, robust, and adaptable emotion recognition systems. These efforts are shaping the way AI systems understand and interact with human emotions, eventually leading to better human-technology interactions across various domains.

The Proposed System

The proposed system goes through different stages, including data preparation, model construction, and evaluation, utilizing advanced techniques such as data augmentation and feature extraction (Xu, 2021; Ullah, 2023).

Dataset

The RAVDESS dataset is a widely used dataset for emotion detection research, developed at Ryerson University, Canada (UWRFKaggler, 2024). It contains 1,440 audio files recorded by 24 professional actors (12 male, 12 female) from diverse ethnic backgrounds, capturing a range of eight emotional expressions: neutral, calm, happy, sad, angry, fearful, disgust and surprised. Each emotion is recorded at three intensity levels (normal, strong, and emphasized) to enhance variability. To help analyze speech characteristics, the dataset includes 162 extracted features, such as Zero Crossing Rate (ZCR): Measures frequency changes in the audio signal, Chroma: Represents pitch classes, MFCC: Captures unique sound frequency patterns, RMS: Measures energy in the signal, and Mel Spectrogram: Visualizes frequency spectrum variations. The audio data is stored in 16-bit, 48 kHz WAV format, with structured file naming conventions that encode details such as modality, vocal channel, emotion type, intensity, and actor ID. In addition to speech, RAVDESS also includes video recordings that capture actors' facial expressions, making it valuable for multimodal emotion recognition research. This dataset is widely used in machine learning, affective computing, and human-computer interaction studies and is publicly available for research purposes through the official website. Its comprehensive emotional diversity and structured format make it essential for developing SER models to enhance AI applications in areas such as chatbots, mental health support, and customer service.

Data Preparation Process

The data preparation process follows a structured approach, as shown in figure 1.

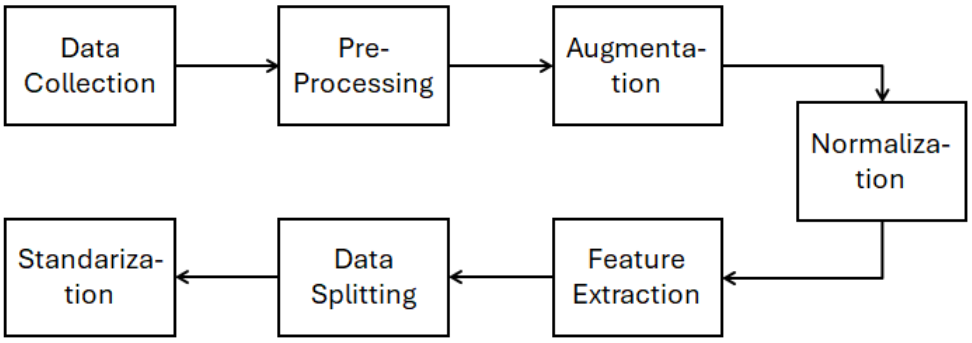

Fig 1. Data preparation flow diagram

1. Data Collection: The RAVDESS dataset is loaded and the file paths with corresponding emotion labels are extracted. Numeric emotion labels are converted into one of the eight descriptive labels (i.e. 'neutral', 'happy', 'sad') for clarity.

2. Data Preprocessing: Emotion labels and file paths are arranged in an understandable structure, enabling efficient analysis and model training. Numeric labels are transformed into their respective emotion names for better readability.

3. Data Augmentation: To enhance the robustness of the model, the augmentation techniques introduce variability, simulating real-world conditions (Jahangir, 2022). Such techniques are Noise Injection: adds random noise to simulate environmental conditions, Speed Variation: alters the speech speed, Time Stretching: adjusts speech timing variations, and Pitch Shifting: modifies vocal pitch.

4. Data Normalization: Emotion labels are one-hot encoded, converting them into a binary matrix representation.

5. Feature Extraction: Essential audio features are extracted to capture key speech characteristics, such as ZCR, Chroma, MFCC, RMS, and Mel Spectrogram. These features are essential for capturing the frequency-based nuances of speech and helping to identify tonal variations linked to emotions.

6. Data Splitting: The dataset is divided into training (80%), validation (10%) and testing (10%) sets to ensure proper model training, fine-tuning, and evaluation.

7. Standardization: Standard scaling is applied to normalize the features. This process involves fitting the scaler to the training data and then transforming both the training and test data. Normalizing the features through standardization is a crucial step in stabilizing the training process. It prevents features with larger scales from disproportionately influencing the model training, ensuring that each feature contributes equally to the learning process.

This structured data preparation process improves the accuracy, generalization and robustness of the SER model, optimizing its ability to recognize and interpret emotions in speech.

Model Architecture

The proposed SER model is built using CNN with regularization techniques to improve accuracy and reduce overfitting. The architecture is shown in figure 2 and consists of the following layers:

- Input layer that matches the shape of input features.

- Three convolutional layers with increasing filters (64, 128 and 256) and a kernel size of 3. Each convolutional layer is followed by batch normalization to stabilize learning, then by Max-Pooling (pool size = 2) to reduce the dimensionality of the features.

- Attention layer focuses on the most relevant parts of the input sequence.

- Fully connected layers, which are three dense layers with different unit sizes (256, 128 and 64). Each layer has ReLu activation and a dropout of 0.5

- Output layer uses softMax activation to classify emotion probabilities based on the eight emotion classes considered in this study.

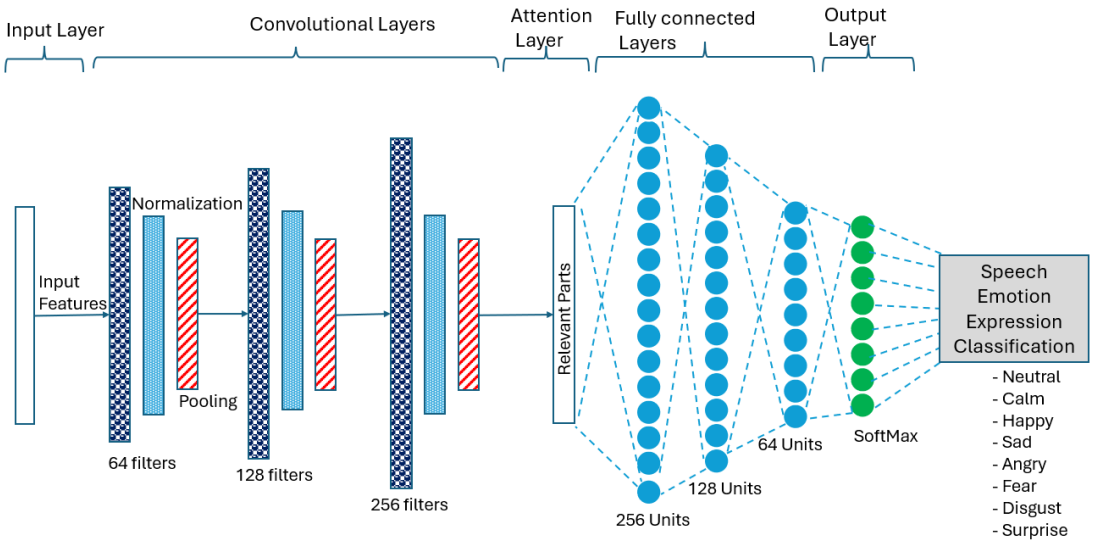

Fig 2. The proposed system architecture using CNN model

The model is compiled using the Adam optimizer, as it outperforms other optimizers (Kingma, 2014). The compilation uses the categorical cross-entropy loss function, which is suitable for multi-class classification tasks. The training uses up to 100 epochs with a batch size of 32 and early stopping to monitor validation loss and stop training when it no longer improves. Training is performed using the training data, with validation on the testing data to monitor performance on unseen data. By incorporating data augmentation, feature extraction, regularization techniques and early stopping, the model aims to achieve better performance and improved generalization in recognizing emotions from speech data.

Results and Discussion

A series of experiments were conducted to evaluate different CNN models for speech emotion recognition. Each experiment involved enhancements such as Batch Normalization, L2 Regularization, Attention Layers, Gated Recurrent Unit (GRU), LSTM and Early Stopping. The models were trained on the RAVDESS dataset, which contains various emotional speech samples (neutral, calm, happy, sad, angry, fearful, disgust, and surprised). Performance was evaluated using test accuracy, loss, training accuracy, and validation accuracy. The training and validation accuracy and loss are plotted over the epochs to visualize the model learning process and generalization capabilities. Table 1 summarizes the results of each experiment.

Table 1: Summary of the experiments conducted using the proposed SER-CNN system

	Experiment	Outcome
1	Baseline CNN Model	- Architecture: Simple CNN with no advanced regularization - Test Accuracy: 78.47% - Observations: Showed steady improvement in accuracy and loss over 50 epochs
2	NN with Batch Normalization and L2 Regularization	- Enhancements: Added Batch Normalization (for stable learning) and L2 Regularization (to reduce overfitting) Test - Accuracy: 82.29% - Observations: Improved generalization, lower validation loss
3	Further Enhanced CNN with Regularization	- Enhancements: Additional Conv1D layers with Batch Normalization and L2 Regularization - Test Accuracy: 82.55% - Observations: Continued accuracy gains, more stable training

4	CNN with GRU Layers	- Enhancements: Added GRU layers to capture temporal dependencies in speech data - Test Accuracy: 72.92% - Observations: GRU layers helped, but overall accuracy slightly dropped due to increased complexity
5	CNN with Further Parameter Tuning	- Enhancements: Adjusted hyperparameters and regularization techniques - Test Accuracy: 73.70% - Observations: Moderate improvement but required additional tuning
6	CNN with Global Average Pooling	- Enhancements: Global Average Pooling to reduce dimensionality and computational load - Test Accuracy: 73.35% - Observations: Balanced learning but lower improvement than expected
7	CNN with Early Stopping	- Enhancements: Early Stopping to prevent overfitting and stabilize training - Test Accuracy: 67.88% - Observations: Early stopping helped avoid overfitting, but model complexity affected accuracy
8	CNN with Bidirectional LSTM and Attention Layer	- Enhancements: Added Bidirectional LSTM (to process speech in both forward and backward directions) and an Attention Layer (to focus on important speech segments) - Test Accuracy: 76.48% - Observations: Improved sequential data understanding, leading to better emotion recognition
9	CNN with Batch Normalization and L2 Regularization	- Enhancements: Further fine-tuned Batch Normalization and L2 Regularization - Test Accuracy: 81.34% - Observations: Strong performance with improved generalization
10	CNN with Custom Attention Layer (Best Model)	- Enhancements: Combined Batch Normalization, L2 Regularization, and a Custom Attention Layer - Test Accuracy: 87.50% (Highest) - Observations: Best model, significantly outperforming previous benchmarks

Experiment 1 is the initial experiment that uses a basic CNN model without any advanced regularization or normalization techniques. It serves as a baseline for measuring the performance improvement achieved in subsequent experiments.

Figures 3 and 4 show the accuracy and loss of the baseline SER-CNN model, respectively. The training accuracy increases steadily until it reaches 1.0 by epoch 50. The validation accuracy follows the trend, but its maximum accuracy is around 0.75. For the training loss, it decreases constantly until it reaches 0.6 by epoch 50. However, the validation loss suffers from high fluctuation, indicating that there is a high variability in the model's performance on the validation set, due to the model encountering different data samples during training. For the other experiments in table 1, regularization techniques (Batch Normalization & L2 Regularization) significantly improved generalization across multiple experiments. Attention mechanisms and bidirectional LSTM improved the ability to focus on relevant emotional speech patterns. Early stopping was effective in preventing overfitting, but excessive computations impacted model performance. The final model (CNN + Custom Attention Layer) achieved the highest accuracy of 87.50%, demonstrating the effectiveness of combining CNN with attention-based mechanisms.

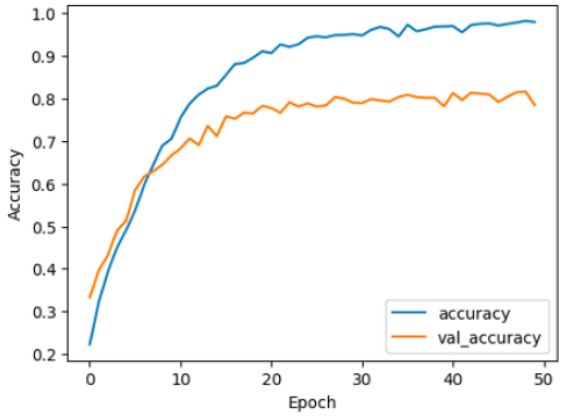

Fig 3. Training and validation accuracy for the baseline model 1

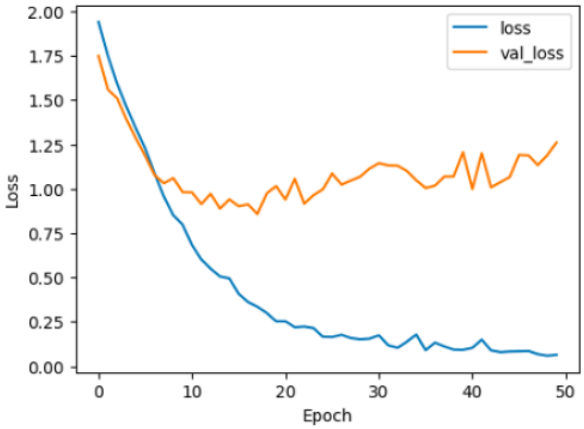

Fig 4. Training and validation loss for the baseline model 1

Figure 5 shows the accuracy of model 10 for the training and validation process over 76 epochs, and figure 6 shows the performance loss of model 10. The accuracy plot shows that the training accuracy starts around 0.8546 and remains high, reaching 0.8750, while the validation accuracy starts around 0.8750 and remains consistent throughout. The loss plot indicates that the training loss starts around 4.8372 and consistently decreases to 0.3782. Similarly, the validation loss starts around 3.0462, fluctuates slightly and ends at 0.3768. The results suggest that the advanced architecture of model 10 with batch normalization, L2 regularization and a custom attention layer achieves better performance and improved generalization with a test accuracy of 87.50%. The test accuracy achieved in this study is a significant improvement compared to previous benchmarks (80.2% by Mustaqeem and Kwon in (Mustaqeem, 2019), and 82.29% by Kapoor and Kumar (Kapoor, 2022)).

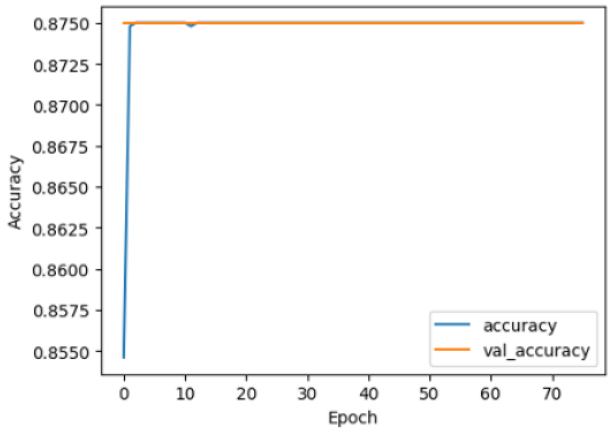

Fig 5. Training and validation accuracy for model 10

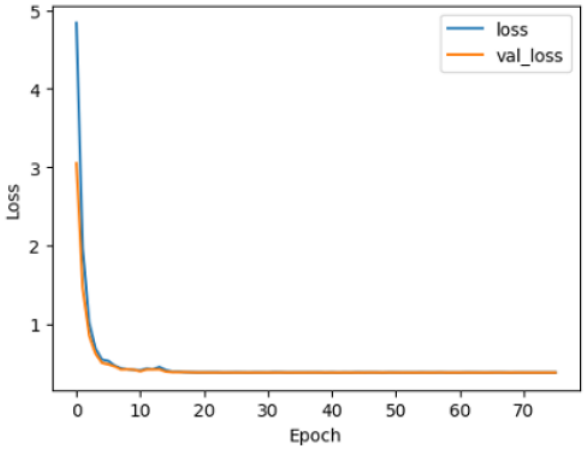

Fig 6. Training and validation loss for model 10

Real-time Voice Emotion Prediction Using the Best Performing Model

The best-performing CNN model from Experiment 10 is implemented in a real-time voice emotion recognition system. This system allows users to record their voice using a microphone, processes the audio in real-time and predicts the emotion using deep learning. The pre-trained CNN model and scaler are loaded to ensure audio data is processed consistently with the training phase. After the user records their voice, the key features for emotion detection (ZCR, Chroma, MFCCs, RMS and Mel spectrogram) are extracted. The extracted features are normalized and reshaped to match the input format required by the CNN model. The pre-processed audio is fed into the CNN model, which outputs a probability distribution over emotion classes. The highest probability class is selected as the predicted emotion, which is one of eight possibilities: neutral, calm, happy, sad, angry, fear, disgust and surprise. The system provides immediate feedback on the detected emotion. This real-time voice emotion recognition system, powered by deep learning, offers practical applications in customer service, mental health support, and AI assistants. By integrating emotion recognition into AI systems, interactions become more engaging, personalized and human-like, significantly improving user experiences.

Conclusion

This research contributes to the accuracy, robustness and real-world implementation of speech emotion recognition (SER) systems. By integrating emotion recognition into AI chatbots, the study paves the way for emotionally artificial intelligence systems that can enhance user experiences in different services such as customer service and mental health. The proposed SER-enhanced AI system utilizes the deep learning CNN model with data augmentation and enhanced model architecture. It addresses key issues in feature extraction and real-world applicability and achieves an accuracy of 87.5%, outperforming previous benchmarks. The techniques introduced in the proposed SER-CNN system contribute to humanizing digital interactions, making them more responsive, intuitive and impactful in everyday life. Data augmentation techniques such as noise injection, speed variation, pitch shifting, and time stretching help to improve the robustness of the model. Additionally, crucial audio features such as Zero Crossing Rate (ZCR), Chroma, Mel-Frequency Cepstral Coefficients (MFCCs), Root Mean Square (RMS) Energy and Mel Spectrogram were extracted to capture the nuances of emotional speech, ensuring more precise emotion detection. Further work can address the evaluation of the SER-enhanced AI system under real-world conditions, such as noisy environments or varied speaker demographics, to give more realistic validation for the model. Future studies could also explore performance improvements through ensemble models or advanced architectures, such as optimized convolutional kernel sizes and innovative pooling techniques. 8 Techniques such as transfer learning and advanced data augmentation also need to be considered to increase model robustness and generalization. One major challenge in SER systems is ethical concerns, particularly regarding privacy and consent. Emotion recognition systems often process sensitive personal data, raising questions about how data is collected, stored, and used. Furthermore, bias in data sets can lead to skewed interpretations, disproportionately affecting certain demographics. Addressing ethical concerns and ensuring user privacy will be critical. Future work may involve developing frameworks to obtain user consent, protect data privacy and ensure transparency in how emotional data is used. Incorporating explainable AI techniques will also be important to help users understand and trust the emotion recognition processes.

References

- Abdelhamid A.A. (2023),'Speech Emotions Recognition for Online Education,' *Fusion: Practice and Applications (FPA),* 10(01), 78-87, doi: https://doi.org/10.54216/FPA.100104
- Abdelhamid A.A. et al. (2022), 'Robust Speech Emotion Recognition Using CNN+LSTM Based on Stochastic Fractal Search Optimization Algorithm,' *IEEE Access*, 10, 49265–49284, doi: 10.1109/ACCESS.2022.3172954.
- Aggarwal A. et al. (2022), 'Two-Way Feature Extraction for Speech Emotion Recognition Using Deep Learning,' *Sensors*, 22(6), doi: 10.3390/s22062378.
- An X. D. and Ruan Z. (2021), 'Speech Emotion Recognition algorithm based on deep learning algorithm fusion of temporal and spatial features,' *J. Phys. Conf. Ser.*, 1861(1), doi: 10.1088/1742-6596/1861/1/012064.
- Bhattacharya S., Borah S., Mishra B.K., and Mondal A. (2022),'Emotion detection from multilingual audio using deep analysis,' *Multimed. Tools Appl.*, 81 (28), 41309–41338, doi: 10.1007/s11042-022-12411-3.
- Choudhary R.R., Meena G., and Mohbey K.K. (2022), 'Speech Emotion Based Sentiment Recognition using Deep Neural Networks,' *J. Phys. Conf. Ser.*, 2236(1), doi: 10.1088/1742-6596/2236/1/012003.
- Dedeoglu M., Zhang J., and Liang R. (2019), 'Emotion Classification Based on Audiovisual Information Fusion Using Deep Learning,' *The 2019 International Conference on Data Mining Workshops (ICDMW), Beijing, China: IEEE*, Nov., 131–134. doi: 10.1109/ICDMW.2019.00029.
- Farooq M., Hussain F., Baloch N. K., Raja F. R., Yu H., and Zikria Y. B., 'Impact of Feature Selection Algorithm on Speech Emotion Recognition Using Deep Convolutional Neural Network,' *Sensors*, 20(21), doi: 10.3390/s20216008.
- Gupta M. and Chandra S. (2021), 'Speech Emotion Recognition Using MFCC and Wide Residual Network,' *The 2021 Thirteenth International Conference on Contemporary Computing (IC3-2021), in IC3 '21. New York, NY, USA: Association for Computing Machinery*, Nov., 320–327. doi: 10.1145/3474124.3474171.

- Hazra S. K., Ema R. R., Galib S. M., Kabir S., and Adnan N. (2022), 'Emotion recognition of human speech using deep learning method and MFCC features,' *Radio electronic Computer Systems*, 4, 161–172, doi: 10.32620/reks.2022.4.13.
- Issa D., Fatih Demirci M., and Yazici A. (2020), 'Speech emotion recognition with deep convolutional neural networks,' *Biomed. Signal Processing Control*, 59, doi: 10.1016/j.bspc.2020.101894.
- Jahangir R., Teh Y. W., Mujtaba G., Alroobaea R., Shaikh Z. H., and Ali I. (2022), 'Convolutional neural network-based cross-corpus speech emotion recognition with data augmentation and features fusion,' *Mach. Vis. Appl.,* 33(3), doi: 10.1007/s00138-022-01294-x.
- Johnson M. (2021), 'The Primacy of Data in Deep Learning NLP for Conversational AI,' *Proceedings of the 30th ACM International Conference on Information Knowledge Management, in CIKM '21. New York, NY, USA: Association for Computing Machinery*, Oct., doi: 10.1145/3459637.3482496.
- Kingma D. P. and Ba J. (2014), 'Adam: A method for stochastic optimization,' *arXiv preprint arXiv:1412.6980*.
- Kapoor S. and Kumar T. (2022), 'Fusing traditionally extracted features with deep learned features from the speech spectrogram for anger and stress detection using convolution neural network,' *Multimed. Tools Appl.,* 81 (21), 31107–31128, doi: 10.1007/s11042-022-12886-0.
- Mustaqeem and Kwon S. (2019), 'A CNN-Assisted Enhanced Audio Signal Processing for Speech Emotion Recognition,' *Sensors*, 20(1), doi: 10.3390/s20010183.
- Naas S.A. and Sigg S. (2020), 'Real-time Emotion Recognition for Sales,' *The 2020 16th International Conference on Mobility, Sensing and Networking (MSN), Tokyo, Japan: IEEE,* Dec., 584–591. doi: 10.1109/MSN50589.2020.00096.
- Pulatov I., Oteniyazov R., Makhmudov F., and Cho Y.I. (2023),'Enhancing Speech Emotion Recognition Using Dual Feature Extraction Encoders,' *Sensors*, 23(14), no. 14, doi: 10.3390/s23146640.
- Singh P., Srivastava R., Rana K.P.S., and Kumar V. (2021),'A multimodal hierarchical approach to speech emotion recognition from audio and text,' *Knowl.- Based Syst.*, 229, doi: 10.1016/j.knosys.2021.107316.
- Sultana S., Iqbal M. Z., Selim M. R., Rashid M. M., and Rahman M. S. (2022), 'Bangla Speech Emotion Recognition and Cross-Lingual Study Using Deep CNN and BLSTM Networks,' *IEEE Access*,.10, 564–578, doi: 10.1109/ACCESS.2021.3136251.
- Tariq Z., Shah S.K., and Lee Y. (2019),'Speech Emotion Detection using IoT based Deep Learning for Health Care,' *2019 IEEE International Conference on Big Data (Big Data), Los Angeles, CA, USA: IEEE*, Dec., 4191–4196. doi: 10.1109/BigData47090.2019.9005638.
- Ullah R. et al. (2023), 'Speech Emotion Recognition Using Convolution Neural Networks and Multi-Head Convolutional Transformer,' *Sensors,* vol. 23(13), doi: 10.3390/s23136212.
- UWRFKaggler, 'RAVDESS Emotional Speech Audio,' Kaggle, [Online], [Retrieved: 15-Jul-2024]. Available: https://www.kaggle.com/datasets/uwrfkaggler/ravdess-emotional- speech-audio.
- Xu M., Zhang F., and Zhang W. (2021), 'Head Fusion: Improving the Accuracy and Robustness of Speech Emotion Recognition on the IEMOCAP and RAVDESS Dataset,' *IEEE Access,* 9, 74539–74549, doi: 10.1109/ACCESS.2021.3067460.