DEVELOPMENT OF A MODEL FOR DETECTING EMOTIONS USING CNN AND LSTM

By

MANISH GOSWAMI * ADITYA PARATE ** NISARGA KAPDE ***

SHASHWAT SINGH **** NITIKSHA GUPTA ***** MEENA SURJUSE ******

*-****** Computer Science and Engineering, S. B. Jain Institute of Technology, Management and Research, Nagpur, India.

https://doi.org/10.26634/jse.19.1.21324

Date Received: 25/10/2024 Date Revised: 30/10/2024 Date Accepted: 04/11/2024

ABSTRACT

This paper presents the development of a real-time deep learning system for emotion recognition using both speech and facial inputs. For speech emotion recognition, three significant datasets: SAVEE, Toronto Emotion Speech Set (TESS), and CREMA-D were utilized, comprising over 75,000 samples that represent a spectrum of emotions: Anger, Sadness, Fear, Disgust, Calm, Happiness, Neutral, and Surprise, mapped to numerical labels from 1 to 8. The system identifies emotions from live speech inputs and pre-recorded audio files using a Long Short-Term Memory (LSTM) network, which is particularly effective for sequential data. The LSTM model, trained on the RAVDEES dataset (7,356 audio files), achieved a training accuracy of 83%. For facial emotion recognition, a Convolutional Neural Network (CNN) architecture was employed, using datasets such as FER2013, CK+, AffectNet, and JAFFE. FER2013 includes over 35,000 labeled images representing seven key emotions, while CK+ provides 593 video sequences for precise emotion classification. By integrating LSTM for speech and CNN for facial emotion recognition, the system shows robust capabilities in identifying and classifying emotions across modalities, enabling comprehensive real-time emotion recognition.

Keywords: Emotion Recognition, Short-Term Memory (LSTM), Convolutional Neural Network (CNN), Recurrent Neural Network (RNN), RAVDEES, CREMA-D, Toronto Emotion Speech Set (TESS), SAVEE, Extreme Learning Machine (ELM), Support Vector Machine (SVM).

INTRODUCTION

Emotions are fundamental in human interactions, helping convey understanding of others' feelings during conversations, whether direct or indirect. However, in digital communication on platforms like WhatsApp and Facebook, emojis are commonly relied upon to convey emotions. When exchanging audio files or communicating through mobile devices, discerning the speaker's emotions becomes challenging. To address this issue, a deep learning model has been developed, capable of recognizing emotions solely from audio inputs. This speech emotion recognition technology holds immense value across sectors such as call centers, entertainment, voice assistance, human-computer interactions, and education systems.

The solution includes a website that accepts sound inputs, enabling real-time audio analysis and prediction of emotions. Recurrent Neural Networks (RNNs) are leveraged for speech emotion recognition, with the model designed to detect specific emotions from audio signals and present them to users through a graphical interface. This innovative approach enhances the understanding of emotions conveyed through audio

i-manager Publications
Member of:

This paper has objectives related to SDG

9 INDUSTRY, INNOVATION AND INFRASTRUCTURE

communication, ultimately improving the user experience.

The workflow for emotion processing and speech recognition typically involves three key stages: selecting the emotional expression database, extracting features, and recognizing emotions.

1. Literature Survey

Kurpukdee et al. (2017) introduced an emotion recognition system that utilized deep learning techniques on both speech and video data. Speech signals were converted into Mel-spectrograms and processed by a convolutional neural network (CNN), while video frames underwent a similar process. The outputs from these CNNs were fused using extreme learning machines (ELMs) and classified using a support vector machine (SVM).

Zhang et al. (2017) developed a real-time system for speech emotion recognition (SER), focusing on continuous speech. Their system incorporated voice activity detection, speech segmentation, signal preprocessing, feature extraction, emotion classification, and statistical analysis. Pantic and Rothkrantz (2000) explored emotion recognition in audio conversations, highlighting its significance in human-machine interaction. They emphasized the challenges of audio emotion analysis due to factors like tone, pitch, and noise. The paper outlined the steps involved in audio-based emotion recognition, including data acquisition, preprocessing, feature extraction, classification, and result analysis.

Schuller et al. (2010) utilized deep and convolutional neural networks (CNNs) for emotion classification based on voice data from the DEAP dataset. They highlighted the applications of SER in human-computer interaction and discussed challenges such as subjective emotions, diverse accents, and speaking styles. Their work leveraged the Librosa library and an MLP classifier to achieve significant accuracy in emotion recognition tasks.

Zhao et al. (2019) reviewed the use of classifiers like k-nearest neighbors (KNN), hidden Markov models (HMM), support vector machines (SVM), artificial neural networks (ANN), and Gaussian mixture models (GMM) for SER. They outlined challenges in the field, such as the variability of speech features and the subjective nature of emotions. Lim et al. (2016) presented a real-time deep learning-based system for emotion recognition using speech input from a PC microphone. Their model, based on long short-term memory (LSTM) networks, achieved high accuracy in recognizing eight basic emotions. They also discussed the architecture, feature extraction using Mel Frequency Cepstral Coefficients (MFCC), and training methods using TensorFlow and Keras.

Zhu et al. (2024) proposed a novel approach to SER, inspired by human brain mechanisms, by designing an implicit emotional attribute classification system. This approach aimed to simulate human emotional perception processes and enhanced emotion recognition by incorporating implicit emotional attribute information into the SER framework. Zisad et al. (2020) developed an SER system focused on detecting emotions in speech from neurologically disordered individuals, facilitating communication. Their system used CNNs and tonal properties for emotion classification, achieving superior performance compared to traditional models.

Koduru et al. (2020) proposed an emotion recognition system based on speech signals, employing feature fusion and classification using SVM and uto-encoders (AE) for feature dimension reduction.Yuan Ingale and Chaudhari (2012) explored deep neural networks for emotion recognition using both audio and video inputs, improving accuracy with advanced architectures. Rao et al. (2013) employed self-supervised learning to enhance emotion recognition models, especially in scenarios with limited labeled data.

Latif et al. (2021) conducted a systematic review of SER research from a machine learning perspective, outlining core challenges and evaluation guidelines. Neumann and Vu (2017) introduced an approach to emotion recognition from audio signals, focusing on acoustic features derived from perceptual evaluation of audio quality (PEAQ). They emphasized features such as perceptual loudness and temporal envelope alterations.

Abbaschian et al. (2021) combined CNNs to analyze both

visual and audio data for emotion recognition, demonstrating that multimodal systems improved recognition accuracy compared to using a single modality. Stolcke et al. (2000) discussed emotion recognition through various modalities such as facial expressions and body language, highlighting SER as a prominent method due to its temporal resolution and cost-effectiveness. Their study explored multiple machine learning models like Random Forest, Multilayer Perceptron, SVM, CNN, and Decision Trees using the RAVDESS dataset for emotion classification.

Michel and El Kaliouby (2003) focused on detecting emotions from audio using machine learning algorithms like KNN, decision trees, and extra-tree classifiers, analyzing acoustic features such as MFCC. Ekman and Keltner (1970) discussed the use of deep learning techniques, particularly MLP classifiers, for SER using the RAVDESS dataset, highlighting SER's importance in human-computer interaction.

2. Emotion Dataset

Achieving a robust and accurate speech emotion recognition system required careful selection and integration of diverse, comprehensive datasets capturing a wide range of emotional expressions. The training and evaluation process included the RAVDESS dataset, renowned for its well-labeled emotional speech samples that cover a wide spectrum of emotions, such as anger, sadness, fear, disgust, happiness, neutrality, and surprise. The SAVEE dataset was also incorporated, providing British English accents and cultural nuances, along with the Toronto Emotion Speech Set (TESS) dataset, offering emotional expressions from actors of various ages and genders. These datasets collectively enriched the training data, enabling the system to generalize effectively across different speech styles, dialects, and emotional contexts, ultimately enhancing the accuracy and reliability of speech emotion recognition capabilities.

2.1 RAVDEES Dataset

The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) is a comprehensive collection of emotional expressions featuring eight distinct emotions portrayed by 24 characters (12 male and 12 female actors). The dataset totals 24.8 GB, with approximately 1.10 GB utilized for this application. It comprises 7,356 audio files that include both spoken sentences and songs. Each emotion is performed at two intensity levels and in both regular speaking and singing voices. Notably, the dataset features a North American English accent, making it a valuable resource for training and evaluating speech emotion recognition systems.

2.2 SAVEE Dataset

The SAVEE database consists of emotional speech recordings from four native English male speakers (DC, JE, JK, and KL), who were postgraduate students at the University of Surrey, aged 27 to 31. The emotional expressions are categorized into distinct classes based on psychological descriptions, including anger, disgust, fear, happiness, sadness, and surprise, aligned with findings from cross-cultural studies by Ekman. The addition of a neutral category brings the total to seven emotion classes, enhancing the dataset's depth.

2.3 CREMA-D Dataset

CREMA-D is a comprehensive dataset comprising 7,442 original clips performed by 91 actors (48 male and 43 female), aged 20 to 74, representing diverse races and ethnicities. The clips include 12 sentences expressed with six emotions: anger, disgust, fear, happiness, neutral, and sadness. Emotion intensity varies among low, medium, high, and unspecified levels, adding depth to the emotional expressions captured. This rich and diverse collection of audio clips makes CREMA-D a valuable resource for studying and developing emotion recognition models.

2.4 TESS Dataset

The Toronto Emotion Speech Set (TESS) consists of 200 target words spoken in the carrier phrase "Say the word _" by two actresses aged 26 and 64. Each actress portrays seven emotions: anger, disgust, fear, happiness, pleasant surprise, sadness, and neutral, resulting in 2,800 audio files. The dataset is organized with recordings sorted into folders by actress and emotion, facilitating easy access and analysis.

2.5 FER2013 (Facial Expression Recognition 2013)

FER2013 is a widely used dataset in facial emotion recognition, created as part of the ICML 2013 Challenges. It contains over 35,000 labeled grayscale images, each with a resolution of 48x48 pixels, depicting human faces categorized into seven emotional states: anger, disgust, fear, happiness, sadness, surprise, and neutral. Due to its size and diversity, FER2013 has become a benchmark for evaluating deep learning models in facial expression recognition. The uniform image format makes this dataset ideal for training Convolutional Neural Networks (CNNs).

2.6 CK+ (Extended Cohn-Kanade Dataset)

The CK+ dataset is a respected resource for facial expression analysis, extending the original Cohn-Kanade dataset. It contains 593 video sequences from 123 subjects, capturing the transition from neutral to peak emotional expressions. The dataset focuses on seven basic emotions: anger, contempt, disgust, fear, happiness, sadness, and surprise. High-quality annotations provided by expert annotators enhance its value. CK+ includes both posed and spontaneous expressions, making it suitable for studying natural emotional behavior over time.

2.7 AffectNet

AffectNet is one of the largest datasets for facial emotion recognition, containing over 1 million facial images collected from the internet. It offers a wide variety of real-world scenarios and conditions, including variations in lighting, pose, background, and occlusions. AffectNet is manually annotated for 11 facial expressions: six basic emotions (happiness, sadness, anger, fear, surprise, disgust), neutral faces, and compound emotions like contempt and disgust. Its quantity and diversity make it a powerful resource for training models for complex emotion recognition tasks.

3. Feature Extraction

In speech emotion recognition, the speech signal contains numerous parameters that reflect emotional characteristics. Selecting the right features for building a model is a challenging task in this domain. Recent research has identified several crucial features for capturing emotional nuances in speech, including energy, pitch, tone, and various spectral features like Linear Prediction Coefficients (LPC), Mel-Frequency Cepstrum Coefficients (MFCC), and modulation spectral features.

Among these, Mel-Frequency Cepstrum Coefficients (MFCC) stand out as a popular choice for feature extraction in emotion recognition tasks. MFCCs capture the spectral characteristics of speech by applying a filter bank that mimics the human auditory system's response, making them effective in capturing both phonetic and emotional information. Additionally, features such as Root Mean Square Error (RMSE) and Zero Crossing Rate (ZCR) are important for extracting emotional features from speech signals. RMSE measures the energy distribution and variations within the signal, while ZCR reflects the signal's oscillation characteristics.

In CNN-based models, feature extraction is performed automatically by the network's convolutional layers, eliminating the need for manual feature engineering. The network learns filters that detect key aspects of facial expressions, such as the curvature of the mouth (indicating a smile) or the raising of eyebrows (indicating surprise). The deep architecture of CNNs allows them to capture both local features (like eye movements) and global patterns (like overall face shape changes).

3.1 MFCC Features

The Mel-Frequency Cepstrum Coefficient (MFCC) is a widely used representation of the spectral properties of voice signals. It is particularly effective for speech recognition because it accounts for human sensitivity to frequency variations. In each frame, the Fourier transform and power spectrum are estimated and then mapped onto the Mel-frequency scale. The process for extracting MFCC features includes several steps: windowing the signal, applying the Discrete Fourier Transform (DFT), taking the logarithm of the magnitude, and warping the frequencies onto the Mel scale. In the study, the LSTM classifier extracted all MFCC features from the dataset, and the resulting feature vectors were used for training the classifier (Vaidya et al., 2023).

3.2 ZCR Features

The Zero Crossing Rate (ZCR) is a fundamental feature in digital signal processing, particularly in speech and audio analysis. It measures the rate at which a signal crosses the zero amplitude threshold. Essentially, ZCR calculates how many times the waveform crosses the horizontal axis (zero amplitude) within a specified time frame, typically measured in milliseconds.The concept of ZCR is rooted in the observation that speech and audio signals exhibit varying levels of oscillation as they change over time. When a signal transitions from positive to negative or vice versa, it crosses the zero axis, indicating a change in polarity or direction. This crossing of zero points is significant as it reflects essential characteristics of the signal, such as pitch, timbre, and rhythmic patterns.

3.3 RMSE Features

Root Mean Square Error (RMSE) is a statistical measure commonly used in feature extraction and analysis, especially in signal processing tasks such as speech emotion recognition. RMSE is crucial for understanding a signal's characteristics, including its energy distribution, variations, and overall quality. It measures the average magnitude of the differences between predicted and actual values in a dataset. It is calculated by taking the square root of the mean of the squared differences between predicted and actual values, providing insight into how well predictions align with ground truth.

3.4 Flow and Motion-Based Features

When dealing with video sequences or dynamic facial expressions, motion-based feature extraction becomes critical. Optical flow is a technique used to track the movement of facial points over time, capturing subtle temporal changes in expressions. This is particularly important for recognizing emotions like surprise or anger, which involve rapid facial movements.Flow and motion-based features play a pivotal role in facial emotion recognition, especially when analyzing video sequences. Unlike static images, videos contain temporal information that captures the progression of facial movements over time, essential for accurately identifying emotions. Traditional static feature extraction methods may fall short in recognizing the subtle, continuous changes that occur in facial muscles.

4. System Architecture and Models

After completing the feature extraction phase, the next step is to create and train the Long Short-Term Memory (LSTM) model, as shown in Figure 1, and then save it for future use. TensorFlow and Keras are employed for training the LSTM network model, particularly because it deals with audio data and benefits from the capabilities of Recurrent Neural Networks (RNNs), which are specifically designed to handle sequential data like speech.

Once the dataset is imported and pre-processed, the model is trained using TensorFlow and Keras. To avoid retraining the dataset each time the model is used, the trained LSTM model is saved using Keras. This is a crucial step, as a well-trained deep learning model is essential for obtaining accurate outputs. The LSTM model is trained for 100 epochs to enhance its accuracy.

The architecture of the LSTM model comprises one hidden LSTM layer with 128 neurons. Following this layer, there are three dense layers, each incorporating continuous dropouts to prevent overfitting. The first dense layer contains 64 neurons with a Rectified Linear Unit (ReLU) activation function, while the second dense layer consists of 32 neurons, also using ReLU as the activation function. Finally, the output layer comprises 8 neurons, representing the 8 different emotion states being classified, and employs the SoftMax activation function.

The system architecture for emotion detection offers users two input options: facial expressions or speech (sound). Upon selection, the workflow begins by processing the chosen input. For facial emotion detection, a camera captures the user's face, and a Convolutional Neural Network (CNN) extracts key facial features to classify emotions. For speech-based emotion detection, the system captures audio through a microphone, extracting features such as Mel-frequency cepstral coefficients (MFCCs) before passing them through a Long Short-Term Memory (LSTM) network to classify emotions. The results from either input are then displayed to the user, showing the detected emotion in real-time. The architecture is

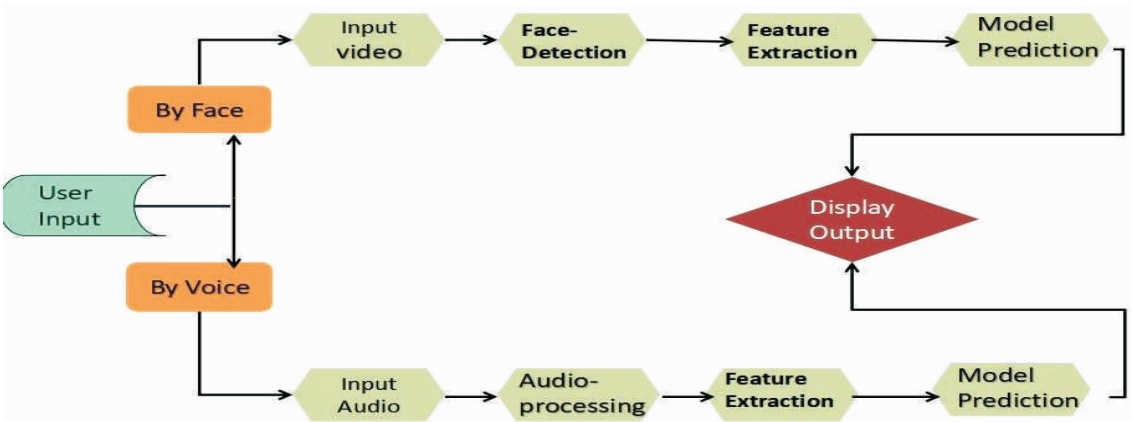

Figure 1. Complete System Architecture

designed to seamlessly switch between face and sound processing based on the user's choice, ensuring a smooth and efficient emotion detection process.

Training a model for emotion recognition in audio data involves several crucial steps, with TensorFlow playing a central role in the process. TensorFlow, an open-source library developed by Google Brain, facilitates numerical computation and large-scale machine learning tasks. It integrates machine learning and deep learning models and algorithms, offering a convenient Python interface while efficiently executing computations in optimized C++. Figure 2 presents a flowchart outlining the emotion recognition process, highlighting the sequence of steps and data transformations involved.

One of the key advantages of TensorFlow is its ability to create a computational graph, where nodes represent mathematical operations and connections represent data flow. This abstraction simplifies the implementation of complex algorithms, allowing developers to focus on the overall logic of the application rather than low-level details. TensorFlow serves as the backend for Keras, a high-level neural networks API, making it a fundamental component in most deep learning tasks.

During training, a crucial element is the loss function, which quantifies model performance by measuring the error between predicted outputs and actual labels in the training and testing sets. This work uses the "Categorical Crossentropy" loss function, suited for multi-class classification tasks, as the dataset includes multiple

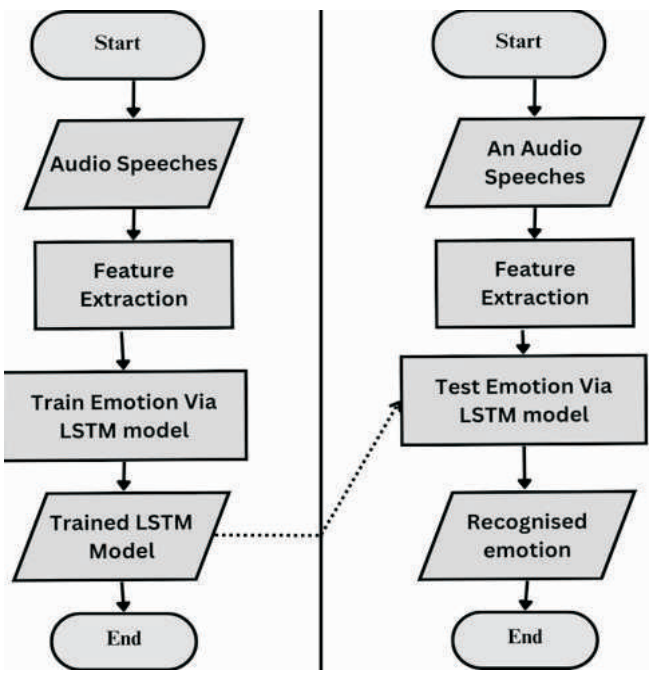

Figure 2. Flowchart

emotion classes. The loss function helps optimize the deep learning algorithm to minimize errors.

The behavior of the loss function is visualized in the Loss Graph, shown in Figure 3, which depicts the decrease in loss over training iterations (epochs). As the model iteratively learns from the data, a declining loss indicates improved accuracy in predicting emotion states. A declining loss value on the Loss Graph not only suggests effective learning but also serves as a diagnostic tool. If the loss stagnates or increases, it may signal potential issues such as overfitting or the need to adjust the learning

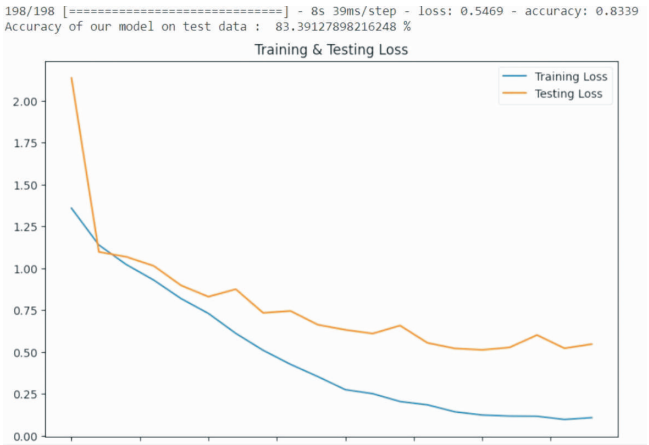

Figure 3. Loss Graph of LSTM Model

rate. Conversely, a steady decrease suggests that training is proceeding effectively. Monitoring the loss function through visualization thus provides essential insights into the training dynamics and the model's performance in recognizing emotions.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

The Categorical Crossentropy loss function is widely used in deep learning, especially for multi-class classification tasks where datasets contain more than two label classes. In this specific implementation, the LSTM model is trained for 20 epochs, with the number of epochs chosen based on desired accuracy and model convergence. During the initial epochs, the model may exhibit higher loss and lower accuracy as it begins learning patterns from the data. However, as training progresses, the loss decreases and accuracy improves, resulting in a well-trained model capable of accurately detecting emotions in audio inputs. Adjusting the number of epochs allows for fine-tuning the model's performance based on requirements and available computational resources.

Since the model is intended to predict human speech emotions, evaluating its performance is crucial before real-time deployment. Evaluation involves assessing the model's reliability and suitability for practical use. In Keras, various evaluation criteria are available; this implementation uses accuracy, as it is effective when all classes in the dataset contain an equal number of records. Given the balanced dataset, accuracy serves as an appropriate measure of performance. The accuracy graph for the CNN model is shown in Figure 4, while the accuracy graph for the LSTM model is shown in Figure 5. These graphs illustrate improvements in model accuracy over training epochs.

The activation functions used in the two Dense layers are 'ReLU' (Rectified Linear Unit) and 'SoftMax'. ReLU introduces non-linearity by activating only positive values, enabling the model to learn complex patterns efficiently. For the output layer, SoftMax is used as it is suitable for multi-class classification, converting raw output scores into probabilities and allowing the identification of the likelihood of each of the eight emotional states targeted for prediction.

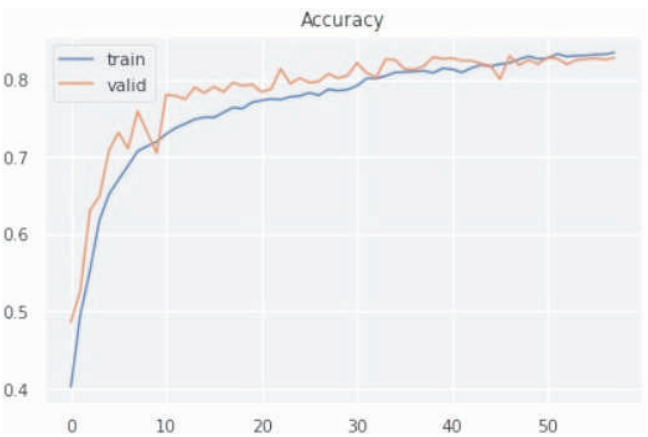

Figure 4. Accuracy Graph of CNN Model

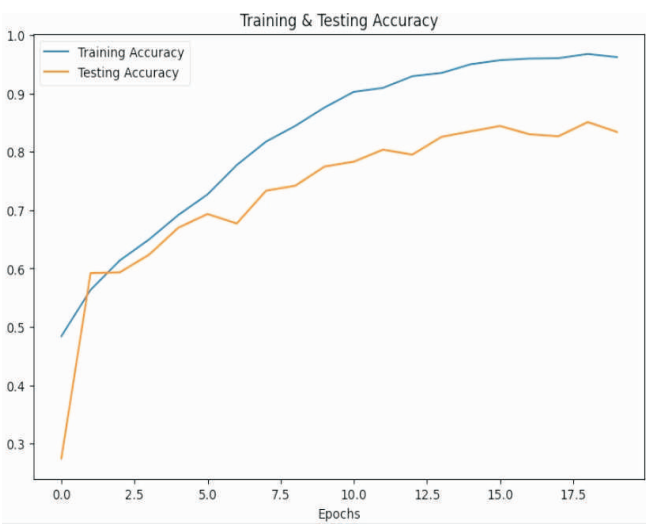

Figure 5. Accuracy Graph of LSTM Model

In the implementation of the CNN-based facial emotion detection model, Categorical Crossentropy was used as the loss function due to its effectiveness in multi-class classification tasks. This loss function measures the difference between the predicted probability distribution and the true label distribution across multiple emotions, penalizing incorrect predictions and encouraging model improvement over time. The model was trained for 100 epochs, allowing it to learn complex features from facial image data. Initially, the model experienced higher loss and lower accuracy as it identified basic patterns, but with continued training, its capability to detect subtle facial expressions improved. This extended training schedule ensured that the model not only converged but also generalized well to new, unseen data.

The ReLU (Rectified Linear Unit) activation function, which is used within the model, plays a significant role in enabling the CNN to learn complex patterns by introducing non-linearity and activating only positive values. Figure 6 shows the ReLU activation function, highlighting how it effectively preserves positive values while filtering out negative values, thereby helping the model efficiently capture essential features for accurate emotion detection.

$$\sigma(\vec{z})_i = \frac{e^{zi}}{\sum_{j=1}^{K} e^{zj}}$$

The architecture of the CNN model includes multiple convolutional layers utilizing the ReLU (Rectified Linear Unit) activation function. ReLU introduces non-linearity by setting negative values to zero while preserving positive values, essential for learning the nuanced patterns of facial expressions. In the final layer, the SoftMax activation function is applied, converting raw output scores into probabilities that sum to 1, enabling the model to accurately predict one of eight possible emotions. The model's performance was assessed using accuracy, an appropriate metric given the balanced distribution of classes in the dataset, as shown in Figure 7. To enhance learning and minimize overfitting, regularization techniques such as dropout layers were incorporated. Figure 8 provides additional metrics, shows the model's efficiency across different performance measures, contributing to robust facial emotion detection. By the end of training, the model exhibited strong performance, accurately detecting and classifying emotions such as anger, fear, happiness, sadness, surprise, and neutral, making it a robust tool for facial emotion recognition in real-world applications (Balvir et al., 2021).

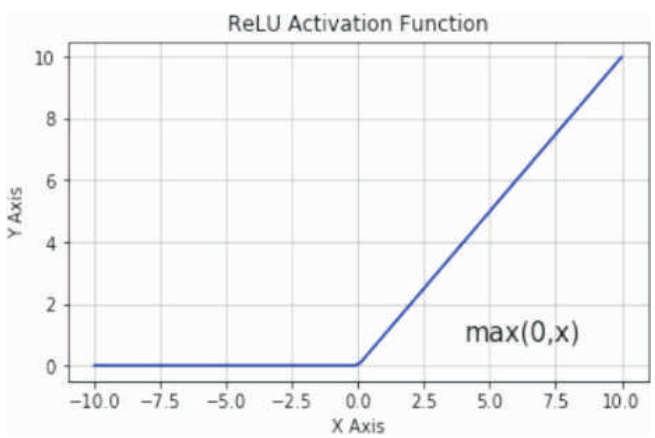

Figure 6. RELU Activation Function Graph

5. Real-Time Emotion Recognition

After completing the training phase, the system proceeds to real-time emotion recognition using audio input from a computer microphone. This involves importing the trained model through Keras to facilitate emotion detection. The Sounddevice module is employed to access the microphone, enabling live audio recording for

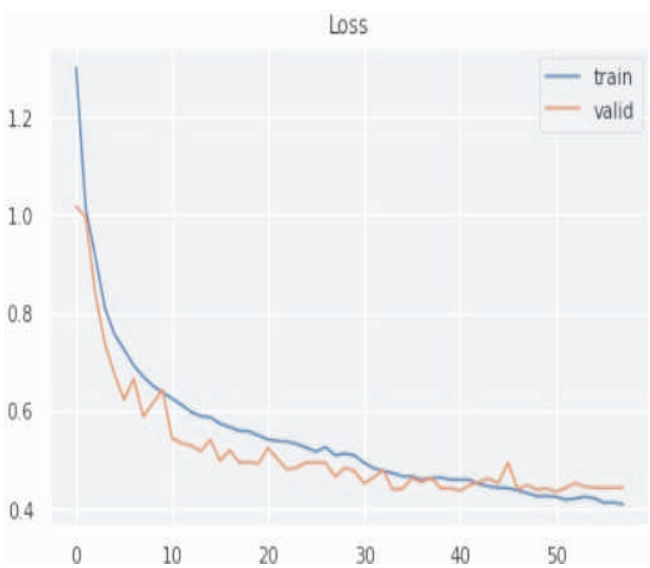

Figure 7. Accuracy of CNN Model

	precision	recall	f1-score	support
angry	0.14	0.13	0.13	3196
disgust	0.01	0.01	0.01	349
fear	0.15	0.10	0.12	3278
happy	0.24	0.25	0.25	5772
neutral	0.18	0.19	0.19	3972
sad	0.17	0.18	0.18	3864
surprise	0.12	0.13	0.12	2537
accuracy			0.18	22968
macro avg	0.14	0.14	0.14	22968
ighted avg	0.17	0.18	0.17	22968

Figure 8. Metrics of CNN model

preprocessing. The system records audio at 44,100 samples per second, although some digital formats support up to 96,000 samples per second. For efficient preprocessing, approximately 48,000 samples per second are processed. During model training, the audio size was set to 48,000 multiplied by 0.8 seconds, which requires resizing the raw audio feed before it is sent to the detection algorithm.

Once preprocessing is complete, the algorithm extracts features such as Mel-Frequency Cepstral Coefficients (MFCCs) and other relevant characteristics from the audio input. The detection algorithm then analyzes these features to determine the speaker's emotion in real time, as shown in Figure 9. This enables the system to seamlessly recognize and interpret emotions during live audio input from the microphone.

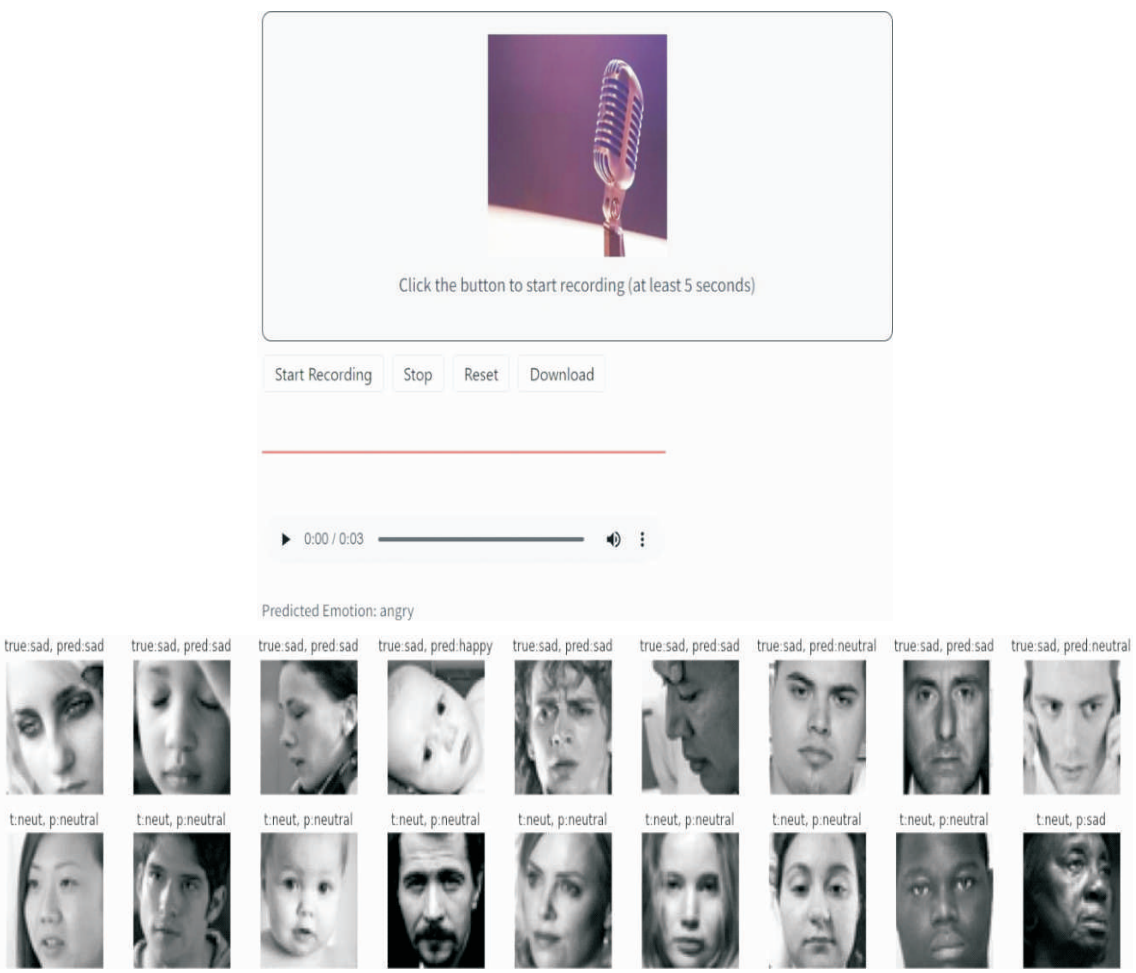

Figure 9. Emotion Recognition of Real Time Audio

6. Results and Discussion

This study introduced a real-time Speech Emotion Recognition system utilizing Neural Networks, specifically Long Short-Term Memory (LSTM) networks for speech and Convolutional Neural Networks (CNNs) for facial emotion recognition. The primary goal of this system is to identify emotions from both speech and facial expressions, particularly beneficial in scenarios such as call centers, where understanding customer emotions is crucial.

6.1 Performance and Accuracy

The system showed strong results in accurately recognizing emotions and associating them with corresponding emojis, enhancing interpretability and user experience. By leveraging the strengths of Neural Networks, particularly RNNs for speech analysis and CNNs for facial recognition, an integrated model was developed that significantly streamlines the emotion detection process. The models outperform traditional methods, capturing temporal dependencies in speech data and spatial hierarchies in facial expressions, which contributes to more accurate predictions.

6.2 Ethical Considerations

Addressing ethical considerations surrounding emotion recognition technology is essential. Issues such as privacy, consent, and the potential for misuse in surveillance or manipulation must be carefully considered. Implementing guidelines and ensuring transparency in how this technology is used will be vital to maintaining trust and safeguarding individuals' rights.

6.3 Practical Implications and Applications

The practical implications of this system extend to various real-world applications, including enhancing user interactions in virtual assistants, improving mental health monitoring, and refining customer service experiences. By integrating emotion recognition into everyday technology, more empathetic and responsive environments can be fostered, adapting to users' emotional states and ultimately improving overall satisfaction and engagement.

Conclusion

In this paper, a real-time Speech Emotion Recognition (SER) system based on deep learning is introduced, specifically leveraging Recurrent Neural Networks (RNNs) such as Long Short-Term Memory (LSTM) networks. This model is designed to identify emotions in audio files or from the speech of specific individuals. High accuracy of 86% was achieved for the training data and 83% for the validation set, demonstrating the system's effectiveness in accurately recognizing emotions. Additionally, the model can recognize emotions in speech files not included in the training data, showcasing its generalization ability.

A Convolutional Neural Network (CNN) was also implemented for facial emotion recognition, achieving 85% accuracy using the FER2013 dataset, which comprises over 35,000 labeled images representing various facial expressions. By combining LSTM for speech emotion detection with CNN for facial emotion recognition, a more comprehensive system capable of understanding emotions from both auditory and visual inputs was created.

This study presents the implementation of Speech Emotion Recognition using deep learning techniques. By enabling machines to discern human emotions through speech and facial expressions, this work contributes to improving communication between humans and machines. Looking ahead, further integration of speech emotion recognition with facial emotion recognition could enhance emotional response accuracy, leading to the development of more sophisticated emotion detection systems. Ultimately, advancements in both speech and facial emotion recognition will pave the way for true artificial intelligence.

References

[1]. Abbaschian, B. J., Sierra-Sosa, D., & Elmaghraby, A. (2021). Deep learning techniques for speech emotion recognition, from databases to models. *Sensors,* 21(4), 1249.

https://doi.org/10.3390/s21041249

[2]. Balvir, S., Sahu, S., & Rohankar, J. (2021). Approaches and applications of sentiment analysis on users data. *Journal of University of Shanghai for Science and Technology,* 23(6), 1761-1767.

[3]. Ekman, P., & Keltner, D. (1970). Universal facial

expressions of emotion. *California Mental Health Research Digest,* 8(4), 151-158.

[4]. Ingale, A. B., & Chaudhari, D. S. (2012). Speech emotion recognition. *International Journal of Soft Computing and Engineering (IJSCE),* 2(1), 235-238.

[5]. Koduru, A., Valiveti, H. B., & Budati, A. K. (2020). Feature extraction algorithms to improve the speech emotion recognition rate. *International Journal of Speech Technology,* 23(1), 45-55.

https://doi.org/10.1007/s10772-020-09672-4

[6]. Kurpukdee, N., Koriyama, T., Kobayashi, T., Kasuriya, S., Wutiwiwatchai, C., & Lamsrichan, P. (2017, December). Speech emotion recognition using convolutional long short-term memory neural network and support vector machines. In *2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)* (pp. 1744-1749). IEEE.

https://doi.org/10.1109/APSIPA.2017.8282315

[7]. Latif, S., Rana, R., Khalifa, S., Jurdak, R., Qadir, J., & Schuller, B. (2021). Survey of deep representation learning for speech emotion recognition. *IEEE Transactions on Affective Computing,* 14(2), 1634-1654.

https://doi.org/10.1109/TAFFC.2021.3114365

[8]. Lim, W., Jang, D., & Lee, T. (2016, December). Speech emotion recognition using convolutional and recurrent neural networks. In *2016 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)* (pp. 1-4). IEEE.

https://doi.org/10.1109/APSIPA.2016.7820699

[9]. Michel, P., & El Kaliouby, R. (2003, November). Real time facial expression recognition in video using support vector machines. In *Proceedings of the 5th International Conference on Multimodal Interfaces* (pp. 258-264).

https://doi.org/10.1145/958432.958479

[10]. Neumann, M., & Vu, N. T. (2017). Attentive convolutional neural network based speech emotion recognition: A study on the impact of input features, signal length, and acted speech. *arXiv preprint arXiv:1706.00612*.

https://doi.org/10.48550/arXiv.1706.00612

[11]. Pantic, M., & Rothkrantz, L. J. M. (2000). Automatic analysis of facial expressions: The state of the art. *IEEE Transactions on Pattern Analysis and Machine Intelligence,* 22(12), 1424-1445.

https://doi.org/10.1109/34.895976

[12]. Rao, K. S., Koolagudi, S. G., & Vempada, R. R. (2013). Emotion recognition from speech using global and local prosodic features. *International Journal of Speech Technology,* 16, 143-160.

https://doi.org/10.1007/s10772-012-9172-2

[13]. Schuller, B., Vlasenko, B., Eyben, F., Wöllmer, M., Stuhlsatz, A., Wendemuth, A., & Rigoll, G. (2010). Cross-corpus acoustic emotion recognition: Variances and strategies. *IEEE Transactions on Affective Computing,* 1(2), 119-131.

https://doi.org/10.1109/T-AFFC.2010.8

[14]. Stolcke, A., Ries, K., Coccaro, N., Shriberg, E., Bates, R., Jurafsky, D., & Meteer, M. (2000). Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational Linguistics,* 26(3), 339-373.

https://doi.org/10.1162/089120100561737

[15]. Vaidya, C. D., Botre, M., Rokde, Y., Kumbhalkar, S., Linge, S., Pitale, S., & Bawne, S. (2023). Unveiling sentiment analysis: A comparative study of LSTM and logistic regression models with XAI insights. *i-manager's Journal on Computer Science,* 11(3).

https://doi.org/10.26634/jcom.11.3.20471

[16]. Zhang, S., Zhang, S., Huang, T., & Gao, W. (2017). Speech emotion recognition using deep convolutional neural network and discriminant temporal pyramid matching. *IEEE Transactions on Multimedia,* 20(6), 1576-1590.

https://doi.org/10.1109/TMM.2017.2766843

[17]. Zhao, J., Mao, X., & Chen, L. (2019). Speech emotion recognition using deep 1D & 2D CNN LSTM networks. *Biomedical Signal Processing and Control,* 47, 312-323.

https://doi.org/10.1016/j.bspc.2018.08.035

[18]. Zhu, X., Huang, Y., Wang, X., & Wang, R. (2024).

Emotion recognition based on brain-like multimodal hierarchical perception. *Multimedia Tools and Applications,* 83(18), 56039-56057.

https://doi.org/10.1007/s11042-023-17347-w

[19]. Zisad, S. N., Hossain, M. S., & Andersson, K. (2020, September). Speech emotion recognition in neurological disorders using convolutional neural network. In *International Conference on Brain Informatics* (pp. 287-296). Springer International Publishing.

https://doi.org/10.1007/978-3-030-59277-6_26

ABOUT THE AUTHORS

Dr. Manish Goswami is an Associate Professor in the Department of Computer Science and Engineering at SB Jain Institute of Technology, Management, and Research, Nagpur, Maharashtra, India. With over 24 years of experience in academia, he specializes in Compiler Design, Theory of Computation, Machine Learning, and Data Science. Holding a Ph.D. in Computer Science and Engineering, he has made significant contributions to these fields and continues to drive research and innovation while imparting his knowledge to students.

Aditya Parate is currently working as an Assistant Professor in the Computer Science Engineering Department at S.B. Jain Institute of Technology, Management, and Research, Nagpur, Maharashtra, India. He is pursuing an M.Tech from Kalinga University, Chhattisgarh, India, and has completed a B.Tech and BCCA. He has experience in Linux and System Administration, Database Development, Web Design, and WordPress Development. He has successfully published two books: Operating System Insights with Linux and The Beginning: Shell Scripting.

Nisarga Kapde is a B.Tech student specializing in Computer Science and Engineering at SB Jain Institute of Technology Management and Research, Nagpur, Maharashtra, India. Aspiring to become a Data Analyst, she has developed a strong foundation in Python, SQL, Machine Learning, Data Analysis, Power BI, and Pandas. With certifications in Data Analytics, Python Data Analytics, Tableau, and Enterprise Data Science, she has worked on projects including the development of a MediPredict system, a Speech Emotion Detection system using voice and facial expressions, and creating interactive dashboards for Olympic and sales data.

Shashwat Singh is a B.Tech Student specializing in Computer Science and Engineering at SB Jain Institute of Technology Management and Research, Nagpur, Maharashtra, India, with a strong foundation in research-oriented data analysis and backend development. His experience includes internships where he conducted exploratory data analysis, improved machine learning model accuracy, and optimized backend systems. Proficient in tools like Spring Boot, Python, and Statistical Modeling, he has developed projects focused on enhancing operational efficiency and data precision.

Nitiksha Gupta is a B.Tech Student specializing in Computer Science and Engineering at SB Jain Institute of Technology, Management, and Research, Nagpur, Maharashtra, India. Aspiring to become a Full-Stack Developer, she has gained valuable experience through internships as a Web Development Intern at Codemate IT Services and as a Data Science Intern at LGPS Pvt. Ltd., Nagpur, India. Skilled in Python, HTML, CSS, and SQL, she has worked on diverse projects, including real-time object detection using YOLO and a speech emotion detection system that leverages both voice and facial expressions.

Meena Surjuse is a B.Tech student specializing in Computer Science and Engineering at SB Jain Institute of Technology Management and Research, Nagpur, Maharashtra, India. Aspiring to become a Full Stack Developer, she has gained experience as a Full Stack Developer Intern at TechVegan Hardware and Software, a Front-End Web Developer at Bharat Intern, and a Data Analyst at LGPS Hybrid Energy Pvt. Ltd., Nagpur, India. She possesses strong technical skills in HTML, CSS, C++, SQL, Python, Machine Learning, and Data Analysis.