Audio Based Speech Emotion Prediction Using CNN Algorithm

Mrs. A.Jeevarathinam1. Ahalya K2
1.Assistant Professor, Department of Computer Science,
2. Student BSc Computer Science
Sri Krishna Arts and Science College, Coimbatore
jeevarathinama@skasc.ac.in,ahalyak22bcs003@skasc.ac.in

ABSTRACT:

Emotion recognition plays a pivotal role in advancing Human-Computer Interaction (HCI) by enabling systems to understand and respond to human emotional states. Among various modalities, Speech Emotion Recognition (SER) stands out as a non-invasive, cost-effective, and temporally efficient approach for detecting emotions. This study leverages the RAVDESS dataset, which provides high-quality audio recordings for emotion classification. The proposed methodology involves preprocessing audio signals to remove noise, extracting temporal and spectral features in both time and frequency domains, and implementing machine learning models for multi-class emotion classification. The study evaluates and compares the performance of several classification models, including Random Forest (RF), Multilayer Perceptron (MLP), Support Vector Machine (SVM), Convolutional Neural Networks (CNN), and Decision Trees (DT). Experimental results demonstrate promising accuracy levels, highlighting the potential of these machine learning techniques in SER. This research contributes to the integration of SER within Brain-Computer Interfaces (BCI) and other emotion-aware applications, paving the way for enhanced interactive systems.

Keywords – Speech Emotion Recognition - Machine Learning - Random Forest -Multilayer Perceptron - Support Vector Machine - Convolutional Neural Networks - Decision Tree - RAVDESS Dataset

1. INTRODUCTION

Emotions play an integral role in human communication, facilitating the exchange of feelings and intentions. They can be expressed through various channels, including facial expressions, body movements, and speech. Among these, speech offers a powerful and accessible medium for emotion recognition due to its unique attributes, such as pitch, tone, loudness, and timbre. Speech Emotion Recognition (SER) focuses on identifying and classifying emotional states through audio signals, enabling seamless interaction between humans and machines.

With the growing importance of Human-Computer Interaction (HCI), emotion recognition has gained significant attention from researchers. It enhances communication in scenarios where face-to-face interaction is not possible, such as virtual meetings, call centers, or assistive technologies. SER also benefits individuals with physical disabilities who may find it challenging to express emotions through traditional means.

This study explores the use of machine learning techniques to classify emotions from speech. Emotions such as happiness, sadness, anger, surprise, fear, and neutrality are universally recognized and form the basis for SER systems. Feature extraction from audio signals plays a critical role in analysing emotions, leveraging temporal and spectral properties of the speech.

Several machine learning models have been widely adopted for emotion classification,each with its unique strengths. Support Vector Machines (SVM) aim to find the optimal hyperplane for classifying data in n-

dimensional space, excelling in both regression and classification tasks. Multilayer Perceptron (MLP), a type of feed-forward neural network, uses backpropagation for training and features input, hidden, and output layers to process data efficiently. Random Forest (RF) is an ensemble learning technique that builds multiple decision trees and uses majority voting to improve accuracy and robustness. Decision Tree (DT) algorithms represent data as tree structures, making classification tasks interpretable and efficient. Convolutional Neural Networks (CNN), a deep learning model, automatically extract relevant features through convolution and pooling operations, offering advanced capabilities for complex emotion recognition tasks.

This paper compares and evaluates these models on the RAVDESS dataset, which provides high-quality speech recordings for emotion classification. The remainder of this paper is organized as follows: Section 2 reviews related work, Section 3 presents the proposed methodology, Section 4 discusses experimental results, and Section 5 concludes the paper with future directions.

2. LITERATURE REVIEW

Speech Emotion Recognition (SER) is a growing field that utilizes machine learning and deep learning techniques to classify human emotions from speech signals. Traditional approaches, such as Support Vector Machines (SVM), k-Nearest Neighbors (k-NN), and Hidden Markov Models (HMMs**)**, relied on handcrafted acoustic features like MFCCs, Spectral Features, and LPC. However, these models faced limitations in handling diverse speakers and real-world noise.

Deep learning models like CNNs, RNNs, LSTMs, and hybrid CNN-LSTM architectures have significantly improved SER accuracy by learning spatial and temporal features from speech data. Recent advancements include attention mechanisms and transformer-based models such as Wav2Vec 2.0 and SpeechT5, which enhance feature extraction and classification.

Researchers have also explored multimodal approaches, combining speech with facial expressions and text for better emotion detection. Popular datasets like IEMOCAP, RAVDESS, and CREMA-D have been widely used for training and evaluation. Despite progress, challenges such as background noise, speaker variability, and class imbalance remain. Ongoing research aims to address these issues through data augmentation, domain adaptation, and self-supervised learning, paving the way for more robust and real-time SER applications in healthcare, human-computer interaction, and affective computing..

Future developments in neuromorphic computing, quantum machine learning, and multimodal affective computing are expected to revolutionize SER, making it more accurate, adaptive, and context-aware. The continuous evolution of synthetic speech and voice cloning technologies also raises concerns about deepfake detection and ethical considerations in SER applications. Addressing these challenges will be crucial in building trustworthy, efficient, and unbiased emotion recognition systems for widespread adoption in real-world scenarios.

Author & year	Methodology	Findings
Zhao et al. (2021)	Deep Convolutional Recurrent Network with LSTM	Achieved promising results using the RECOLA dataset for automatic speech emotion recognition.
Kim & Lee (2020)	1D CNN-LSTM and 2D CNN-LSTM on IEMOCAP dataset	Demonstrated the effectiveness of combining temporal and spatial features for emotion recognition.
Wang et al. (2019)]	Deep Neural Network (DNN) on silent video data	Reported 70% accuracy for visual emotion detection and 75% for voice-based analysis.
Patel & Singh (2018)	Support Vector Machine (SVM) on LDC and UGA datasets	Focused on gender-based emotion classification, highlighting variations in emotional patterns.
Chen et al. (2017)	Deep Convolutional Neural Network (DCNN)	Speaker-dependent model achieved 76.96% accuracy, while speaker-independent accuracy was 65.32%.
Haque & Roy (2016)	Comparison of DNN and SVM using IEMOCAP dataset	Models achieved limited accuracy, falling below 55%.

3.PROPOSED WORK

This section outlines the proposed methodology for Speech Emotion Recognition (SER) and is represented schematically in Fig. 1. The approach employs the RAVDESS dataset as input for experimentation due to its high-quality recordings and balanced representation of emotions.

3.1 DATASET DESCRIPTIONS

The below datasets contain speech and song recordings from actors expressing various emotions. They include crowd-validated and controlled environment recordings, ensuring accuracy in emotional expression. Some focus on male or female voices, while others provide multimodal data, including audio and video.

3.1.1 Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS)

The RAVDESS dataset is a high-quality emotional speech and song dataset widely used for SER. It contains recordings from 24 professional actors (12 male and 12 female), producing a total of 7,356 files. Each recording represents one of the following emotions: happiness, calm, sadness, anger, surprise, disgust, and fear. Actors vocalize two statements, “Kids are talking by the door” and “Dogs are sitting by the door,” with varying emotional intensities (normal and strong). The dataset includes audio-only, video-only, and audio-video modalities. For this project, only the audio files are utilized. The recordings maintain consistent audio quality, making them suitable for machine learning applications. This dataset is widely used in emotion recognition research and speech analysis studies.

	Emotions	Path
0	surprise	/kaggle/input/ravdess-emotional-speech-audio/a...
1	neutral	/kaggle/input/ravdess-emotional-speech-audio/a...
2	disgust	/kaggle/input/ravdess-emotional-speech-audio/a...
3	disgust	/kaggle/input/ravdess-emotional-speech-audio/a...
4	neutral	/kaggle/input/ravdess-emotional-speech-audio/a...

	Emotions	Path
1435	fear	/kaggle/input/ravdess-emotional-speech-audio/a...
1436	angry	/kaggle/input/ravdess-emotional-speech-audio/a...
1437	sad	/kaggle/input/ravdess-emotional-speech-audio/a...
1438	disgust	/kaggle/input/ravdess-emotional-speech-audio/a...
1439	angry	/kaggle/input/ravdess-emotional-speech-audio/a...

neutral	288
surprise	192
disgust	192
fear	192
sad	192
happy	192
angry	192

Name: Emotions, dtype: int64

Fig 1. Details of the Ravdess Dataset

3.1.2 Crowd-Sourced Emotional Multimodal Actors Dataset (CREMA-D)

The CREMA-D dataset includes audio-visual recordings of emotional speech from 91 actors (48 male and 43 female), aged between 20 and 74 years, from diverse ethnic backgrounds. It consists of 7,442 recordings, where actors deliver 12 sentences designed to elicit six emotions: happiness, sadness, anger, fear, disgust, and neutrality. The dataset also includes varying emotional intensities and diverse accents, making it ideal for robust SER model training.

disgust	1271
happy	1271
sad	1271
fear	1271
angry	1271
neutral	1087

Name: Emotions, dtype: int64

Fig 2. Details of the Crema-dDataset

3.1.3 Surrey Audio-Visual Expressed Emotion (SAVEE)

The SAVEE dataset comprises recordings from four male speakers expressing seven different emotions: anger, disgust, fear, happiness, sadness, surprise, and neutrality. The dataset contains a total of 480 utterances, with each speaker articulating 15 sentences per emotion. SAVEE is particularly useful for studying audio-visual emotion recognition systems, although this project focuses on the audio modality.

neutral	120
happy	60
fear	60
disgust	60
angry	60
surprise	60
sad	60

Name: Emotions, dtype: int64

Fig 3. Details of theSaveeDataset

3.1.4 Toronto Emotional Speech Set (TESS)

The TESS dataset is designed to explore emotional perception in older and younger adults. It consists of recordings from two female speakers aged 26 and 64, expressing seven emotions: anger, disgust, fear, happiness, sadness, surprise, and neutrality. Each speaker delivers 200 target words in a neutral carrier phrase, such as “Say the word ___.” The dataset includes 2,800 audio files, making it a compact yet valuable resource for SER research.

fear	400
angry	400
disgust	400
neutral	400
sad	400
surprise	400
happy	400

Name: Emotions, dtype: int64

Fig 4. Details of the TessDataset

3.2 DATA PREPROCESSING

In the preprocessing stage, data augmentation techniques are applied to artificially expand the dataset. These techniques involve generating new data by modifying existing samples, which helps

improve the model's performance and generalization. One common method of augmentation includes injecting noise into audio signals to evaluate the model's robustness against variations. Additionally, the generation of synthetic data for Mel-Frequency Cepstral Coefficients (MFCC) has proven effective in enhancing the recognition of speakers using transfer learning approaches. These steps are critical for improving the quality and diversity of the training data.

3.3 FEATURE SELECTION AND EXTRACTION

Feature extraction is a fundamental step in audio classification, as it helps in isolating meaningful characteristics from raw audio signals. This approach focuses on five key features:

Chroma features analyze the pitch content of an audio signal, categorizing it into 12 semitones of a musical octave. MFCCs (Mel Frequency Cepstral Coefficients) are widely used in speech processing as they capture vocal tract characteristics and consist of 39 features for speaker and speech recognition. The Mel-Spectrum represents the short-term power spectrum of the signal, providing insights into its frequency distribution. Contrast enhances modulation by highlighting differences between strong and weak energy bands, while Tonnetz helps refine tonal qualities, improving audio classification.

A total of 120 features are extracted, with these five being the most prominent for the proposed method. The dataset is divided into training, validation, and testing sets with a 70:20:10 split ratio. Data standardization isperformed to eliminate biases caused by variations in attribute scales, ensuring all features contribute equally during training.

3.4 CLASSIFICATION

Classification involves training models to recognize emotions from extracted speech features. Various deep learning models, including Long Short-Term Memory (LSTM), Convolutional Neural Networks (CNN), and Convolutional Long Short-Term Memory (CLSTM), are implemented to analyze their effectiveness in Speech Emotion Recognition (SER). Additionally, a confusion matrix is used to evaluate the models' performance based on classification accuracy and misclassification patterns.

3.4.1 LSTM MODEL

LSTM is a type of Recurrent Neural Network (RNN) specifically designed to process sequential data by capturing long-term dependencies. It is particularly effective in SER as it retains contextual information from speech signals over time. In this study, the LSTM model is trained using an 80:20 train-validation split. To enhance generalization, batch normalization and dropout techniques are incorporated to mitigate overfitting. The architecture consists of multiple LSTM layers followed by dense layers that output emotion classifications. By leveraging memory cells and gating mechanisms, LSTM efficiently identifies patterns in speech data, making it a reliable choice for emotion recognition. Additionally, activation functions like ReLU and softmax are used to improve learning efficiency and classification performance. Hyperparameter tuning is also performed to optimize model accuracy and reduce training loss.

3.4.2 CNN MODEL

CNN, a widely used deep learning model, is employed to extract spatial features from speech

spectrograms and perform classification. Unlike traditional machine learning approaches, CNN automatically learns hierarchical representations of features through convolutional operations. In this study, multiple CNN architectures with two, three, and four convolutional layers are explored to evaluate their impact on classification performance. The architecture comprises convolutional layers activated by ReLU functions, max-pooling layers to reduce dimensionality, dropout layers to prevent overfitting, and fully connected layers for final classification. By capturing local features within speech signals, CNN enhances the ability to distinguish between different emotional states.

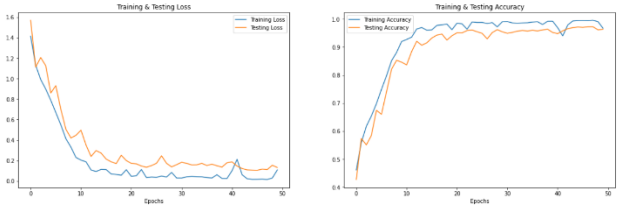

Fig 5. Details of CNN Model

3.4.3 CLSTM MODEL

A hybrid approach, CLSTM, is implemented to combine the strengths of CNN and LSTM models. CNN extracts spatial features from spectrogram representations of speech data, while LSTM captures temporal dependencies, enabling a comprehensive understanding of emotion patterns. This integration allows the model to utilize both spatial and sequential information, improving classification accuracy.

The CLSTM model is trained using an Adam optimizer, and batch normalization and dropout techniques are applied to ensure robust performance. By leveraging the complementary advantages of both CNN and LSTM, the CLSTM model aims to improve emotion classification results. This hybrid approach allows the model to effectively capture both spatial and temporal features of speech signals, enhancing its ability to recognize emotions accurately.

	precision	recall	f1-score	support
angry	0.96	0.97	0.97	1484
disgust	0.97	0.95	0.96	1558
fear	0.96	0.97	0.96	1505
happy	0.96	0.95	0.96	1619
neutral	0.97	0.98	0.97	1558
sad	0.96	0.97	0.96	1478
surprise	0.98	0.97	0.97	528
accuracy			0.96	9730
macro avg	0.96	0.96	0.96	9730
weighted avg	0.96	0.96	0.96	9730

Fig 6. Details of CLSTM Model

3.4.4 CONFUSION MATRIX

To assess the effectiveness of these models, a confusion matrix is generated for each classification method. The confusion matrix provides a detailed evaluation of correctly and incorrectly classified emotions, helping to analyze misclassification trends. Metrics such as accuracy, precision, recall, and F1-score are derived from the confusion matrix to measure the models' overall performance. A comparative analysis of confusion matrices across different models helps identify challenges in emotion recognition and provides insights into potential improvements. By evaluating classification results through these metrics, the study aims to determine the most effective model for SER while balancing computational efficiency and accuracy. Additionally, analyzing false positives and false negatives helps in understanding emotion overlap and model biases. Further, optimizing feature selection and model parameters can enhance classification performance.

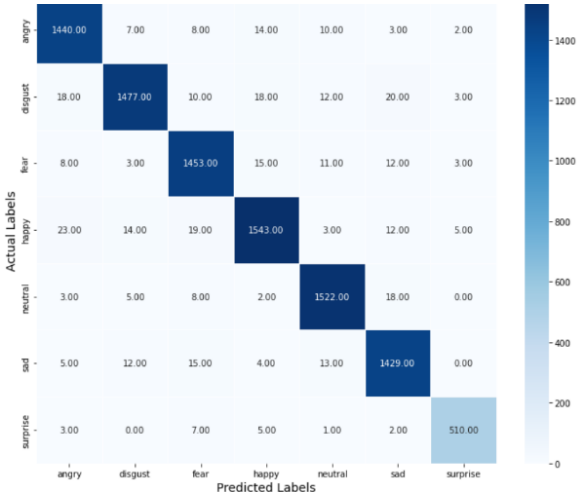

Fig 7. Details of Confusion Matrix

4. RESULT & DISCUSSION

The Speech Emotion Recognition (SER) system effectively classifies emotions from audio input, providing an intuitive and interactive user experience. The project features an animated start page that guides users seamlessly into the detection process. Upon selecting an audio file through the "Choose File" option and clicking "Predict Emotion," the system processes the input and accurately determines the dominant emotion, such as happy, sad, angry, fear, or neutral. Additionally, if the model's confidence is below 100%, the system presents alternative possible predictions with corresponding probabilities, ensuring transparency in classification. To enhance interpretability, the system also visualizes the analysis using **pie** charts, bar graphs, and other graphical representations, making the output more comprehensible for users. The integration of an engaging UI with real-time analysis demonstrates the effectiveness of the proposed model in recognizing human emotions from speech.

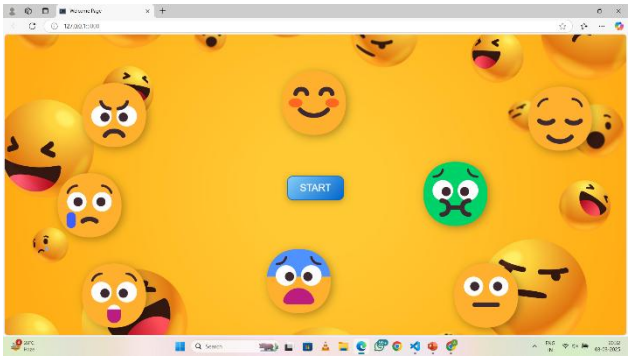

Fig 8. Welcome page

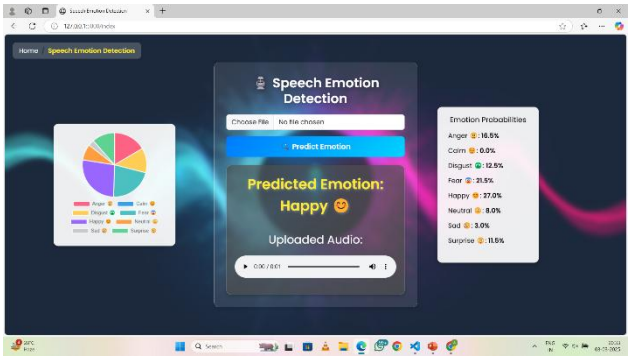

Fig 9. Result with other possibilities

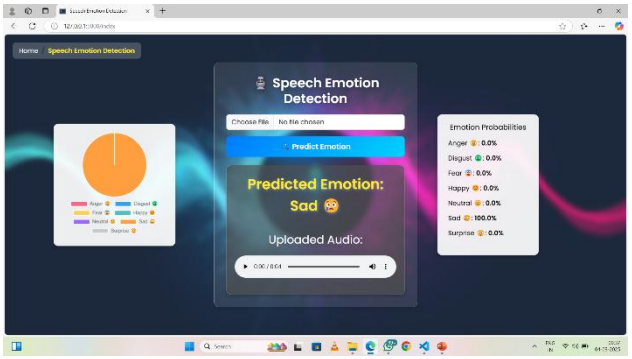

Fig 10. Result with 100% accuracy

5. CONCLUSION & FUTURE WORK

This paper presents a novel approach for emotion classification using audio signals. The proposed system utilizes five machine learning models—SVM, Decision Tree, Random Forest, MLP, and CNN—to classify emotions based on five key features extracted from audio data. The performance of these models was evaluated on a

dataset with emotions such as happy, sad, and neutral. The results showed that the Random Forest model achieved the highest accuracy at 85.71%, followed by CNN at 82.98%, MLP at 81.82%, SVM at 78.57%, and Decision Tree at 78.56%.

The findings indicate that the proposed approach demonstrates promising results in emotion classification using audio signals. Future work will focus on extending the model to classify compound emotions, such as happily surprised, happily disgusted, sadly fearful, sadly angry, sadly surprised,sadly disgusted, and angrily fearful. Additionally, a comparison of various machine learning algorithms on these compound emotions will be explored to enhance the system's performance further.

6. REFERENCES

[1]Clustering-based speech emotion recognition by incorporating learned features and deep BiLSTM. IEEE Access, 8, 79861-79875.

[2] A. Krizhevs[1] C. Peng, "Research emotional and recognition in speech signal," Journal of Jing Hinton, "Image net classification with deep neural networks," in Proceedings of the 26th Annual Conference on Neutral Information Processing Systems (NIPS '12), pp. 1097–1105, Lake New, UK, December 2012.

[3]L. Zhao, C. Jiang, C. Zou, and Z. Wu, "Study on emotional recognisation analysis in speech," Acta Electronica volve. 32, no. 4, pp. 606–609, 2004.

[4]G. Penguan "Research on emotional recognition," Application , vol. 24, no. 10, pp. 101–103, 2007.

[5]L. Zhao, X. Qian, C. Zhou, and Z. Wu, "Study on emotional feature discover from speech ," Journal of Data and Processing, vol. 15, no. 1, pp. 120–123, 2000

[6] Tzirakis, P., Zhang, J., & Schuller, B. W. (2018). End-to-end speech emotion recognition using a deep convolutional recurrent network. ICASSP. 2. Zhao, J., Mao, X., & Chen, L. (2018). Learning deep features to recognise speech emotion using merged deep CNN. IET Signal Processing, 12(6), 713-721..

[7]Real-Time Fraud Detection Systems for Telecom Networks: A Machine Learning Approach by P. Sharma and R. Kumar (2021)

[8]Mirsamadi, S., Barsoum, E., & Zhang, C. (2017, March). Automatic speech emotion recognition using recurrent neural networks with local attention. In 2017 IEEE International conference on acoustics, speech and signal processing (ICASSP) (pp. 2227–2231). IEEE.

[9] Aouani, H., & Ayed, Y. B. (2020). Speech emotion recognition with deep learning. Procedia Computer Science, 176, 251-260. 10. Sajjad, M., & Kwon, S. (2020).