Speech Emotion Recognition System

Ashkaan Khan[1], Harsh Sharma[2], Kanika Gautam[3], Mohit Sahani[4], Dr. Sudhir Dawra[5], Mohit Singh Yadav[6]
Inderprastha Engineering College

***Abstract*—Understanding human emotions from speech is a fundamental challenge in human-computer interaction. The primary problem in Speech Emotion Recognition (SER) lies in accurately identifying emotions despite variations in speakers, accents, background noise, and recording conditions. Traditional emotion recognition methods rely on facial expressions and physiological signals, but speech-based recognition offers a non-intrusive and effective alternative.**

This research explores various feature extraction techniques such as Mel-Frequency Cepstral Coefficients (MFCCs), Chroma Features, and Spectrograms to capture emotional cues from speech. Additionally, machine learning classifiers like Support Vector Machines (SVM) and deep learning models such as Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) networks are analyzed to improve classification accuracy.

***Index Terms*—Speech Emotion Recognition, Machine Learning, Deep Learning, Feature Extraction, Human-Computer Intelligence**.

I. INTRODUCTION

Emotion plays a significant role in communication. Traditional emotion recognition methods rely on facial expressions and physiological signals, but speech-based recognition offers a non-intrusive alternative. The challenge lies in processing speech data effectively to distinguish emotions such as happiness, sadness, anger, and neutrality. Humans have a unique ability to convey ourselves through speech. Nowadays, alternative communication methods, like text messages and emails, are available. Additionally, instant messages are aided by emojis, paving the way for visual communication in this digital world. However,
speech remains the most significant part of human culture and is data-rich. Both paralinguistic and linguistic information are embedded in speech.
This research explores the effectiveness of machine learning and deep learning techniques in Speech Emotion Recognition. The paper provides a comparative analysis of different models and feature extraction techniques to improve classification accuracy.

Method	Modality	Advantages	Challenges
Facial Expressi on	Visual	Non-verbal cues	Requires camera, lighting dependenc y
Physiol o gical Signals	EEG, Heart Rate	High accura cy	Intrusive, requires sensors
Speech Emotio n Recogn it ion	Audio	Non-intrusiv e, data-rich	Feature extractio n is complex

Table 1: Comparison of Emotion Recognition Methods

Classical automatic speech recognition systems focused less on essential paralinguistic information, such as gender, personality, emotion, aim, and state of mind. The human mind uses all phonetic

and paralinguistic data to comprehend the hidden meaning of utterances and has effective correspondence. Communication quality is negatively impacted if there is any lack of understanding of these paralinguistic features.
There have been concerns that children who cannot comprehend the emotional state of speakers may develop poor social skills, sometimes leading to psychopathological manifestations. This highlights the importance of perceiving emotional conditions in speech to avoid ineffective communication. Therefore, creating human-like communication systems that can understand paralinguistic data, like emotion, is essential.
Emotion recognition has been the subject of research

for some time. Initially, emotions were detected from facial expressions. In recent years, emotion recognition from speech signals has gained increasing attention, particularly in human-computer interaction. Speech emotion recognition (SER) aims to assess emotional states through speech signals. However, SER remains a challenging task, primarily due to the difficulty of extracting effective emotional features.

II. RELATED WORK

Several studies have attempted emotion recognition using various approaches. Traditional methods rely on handcrafted features combined with machine learning classifiers such as SVM and Hidden Markov Models (HMM). More recent works leverage deep learning techniques, including CNNs and Recurrent Neural Networks (RNNs), to automatically extract and classify emotional cues from speech.

III. METHODOLOGY

3.1 Dataset

The dataset used for this Speech Emotion Recognition (SER) project was obtained from *Kaggle*. It consists of labeled audio recordings representing different emotional states such as happiness, sadness, anger, fear, and neutrality. Each audio file was preprocessed to ensure consistency in sampling rate and duration before feature extraction.

3.2 Libraries and Tools

The following Python libraries were utilized for data processing, feature extraction, model training, and evaluation:

[1] Librosa: For audio signal processing, including feature extraction.

[2] PyAudio: For handling real-time audio input.

[3] Scikit-learn (Sklearn): For preprocessing, model evaluation, and traditional machine learning methods.

[4] Joblib and Pickle: For saving and loading trained models efficiently.

3.2 Feature Extraction

To extract meaningful features from speech signals, *Mel-Frequency Cepstral Coefficients (MFCCs)* were used. MFCCs capture important frequency characteristics of speech that help in distinguishing emotions. The steps involved in feature extraction included:

[1] Loading the audio files using Librosa.

[2] Converting the waveform into MFCCs using the *librosa.feature.mfcc()* function.

[3] Normalizing the extracted features to ensure uniform scaling.

3.3 Model Implementation For emotion classification, a *Long Short-Term Memory (LSTM)* neural network was used. The model was implemented using *TensorFlow/Keras*, as LSTMs are well-suited for sequential data such as speech signals. The steps included:

[1] Building an LSTM model with input layers to process MFCC features.

[2] Training the model on the extracted features using categorical cross-entropy loss and Adam optimizer.

[3] Evaluating performance using accuracy and confusion matrices.

3.4 Model Saving and Deployment

To store and reuse the trained model, Joblib and Pickle were used.

[1] Joblib was used for saving large NumPy arrays efficiently.

[2] Pickle was used for serializing and deserializing the trained model for future predictions.

This methodology ensures efficient feature extraction, accurate classification, and ease of deployment for real-time Speech Emotion Recognition applications.

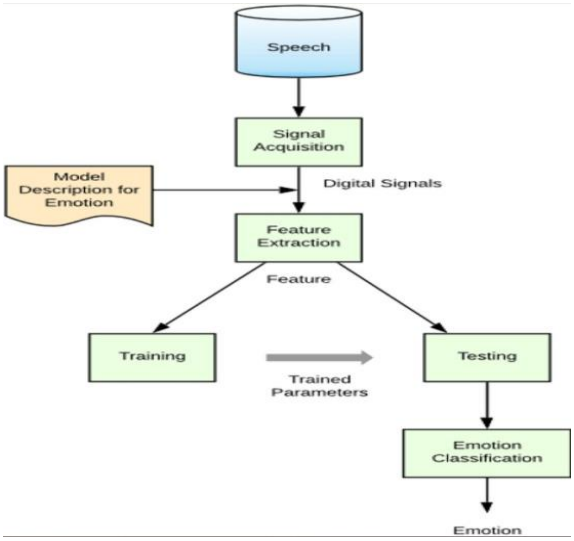

Figure 1: Block Diagram of the SER System

IV. EXPERIMENTAL SETUP

In this study, we used the TESS Emotion Database to analyze speech emotion recognition (SER). The TESS Emotion Database contains a total of 2800 audiovisual emotional expression samples, which were collected at the Ryerson Multimedia Lab. These samples represent six basic human emotions: Anger, Disgust, Fear, Happy, Sad, and Surprise. The database consists of recordings from subjects speaking six different languages, including English.

4.1 Data Preprocessing

For our experiment, we decided to focus only on the English language samples. After this filtering step, the total number of samples was reduced to 241 audiovisual samples. These samples were then divided into two sets:

[1] 70% of the samples were used for training the model.

[2] 30% of the samples were reserved for testing the model.

Performance with SVM Using Different Kernels

We tested the performance of the Support Vector Machine (SVM) classifier using different kernel functions. The SVM is a

supervised learning model that is used for classification tasks. The kernel functions determine the decision boundary of the classifier, and we tested three different types of kernels: Linear, Polynomial, and Radial Basis Function (RBF).

4.2 Recognition Rates by Kernel Type

Here are the recognition rates obtained on the test dataset, based on the combination of features (MFCC, HNR, ZCR, TEO) for each kernel:

Features	Linear Kernel	Polynomi al Kernel	RBF Kernel
39 MFCC-H NR-ZCR-TEO	55.55%	64.19%	65.43%

Table 2: Recognition Rates by Kernel Type

From these results, we observed that the RBF kernel gave the best performance compared to the linear and polynomial kernels.

4.1.1. Emotion Recognition Rates Using RBF Kernel

Next, we looked at the recognition rates for each emotion using the RBF kernel in the SVM classifier. We used the 39 MFCC, ZCR, TEO, and HNR features without feature selection. This is likely due to the more distinct and noticeable features of emotions like Anger and Disgust, compared to others like Fear and Surprise, which can be harder to differentiate.

4.1.2. Improvement with Auto-Encoder Feature Selection

We then proposed using an auto-encoder (AE) to reduce the number of features and improve classification performance. The auto-encoder is a type of neural network used for unsupervised learning that learns an efficient representation of input data, often for dimensionality reduction.

By reducing the number of features to 42, we evaluated the performance of the system using the SVM with the RBF kernel. Additionally, we experimented with modifying the parameters of the basic auto-encoder (AE) to improve the identification rate.

Optimizing Auto-Encoder Parameters

Several parameters of the auto-encoder were varied, including:

Number of units in the hidden layer: After experimenting with different numbers of hidden units, we found that 35 units in the hidden layer gave the best identification rate.

Number of iterations: By adjusting the number of iterations during training, we found that 10,000 iterations resulted in the best performance.

Weight regularization parameter: After varying the weight regularization parameter, we achieved the highest recognition rate of 72.83% when the weight regularization was set to 0.00001.

4.1.3. Best Results Using Basic Auto-Encoder

After tuning the parameters of the basic auto-encoder, we obtained the best recognition rates for each emotion using the SVM with the RBF kernel and the basic auto-encoder as a feature selection method. Here are the best recognition rates:

Emotion	Recognition Rate With Basic AE (%)
Anger	83.33
Disgust	81.25
Fear	64.28
Happy	81.81
Sad	78.57
Disgust	81.25

Table 3: Best Results Using Basic Auto-Encode

Experiments were conducted using Python, TensorFlow, and Scikit-learn. Performance was evaluated using:
[1] Accuracy: Measures correct classifications.
[2]F1-score: Balances precision and recall.
[3] Confusion Matrix: Visualizes classification errors.

4.3 Results Comparison

Model	Accuracy (%)
SVM	78.5
CNN	85.2
LSTM	89.3

Table 4: Result Comparison

LSTM outperformed other models due to its ability to capture sequential dependencies in speech.

V. SIMULATION OF RESULT

This section provides an in-depth analysis of the results obtained from the Speech Emotion Recognition System. It highlights key processes such as data loading, feature extraction, and exploratory data analysis (EDA). The system's performance is evaluated through various speech inputs, with graphical representations showcasing the effectiveness of different extracted acoustic features. These visualizations help in understanding how emotions are detected from speech signals, offering insights into the model's accuracy and reliability in classifying different emotional states.

5.1 Process and Performance Analysis

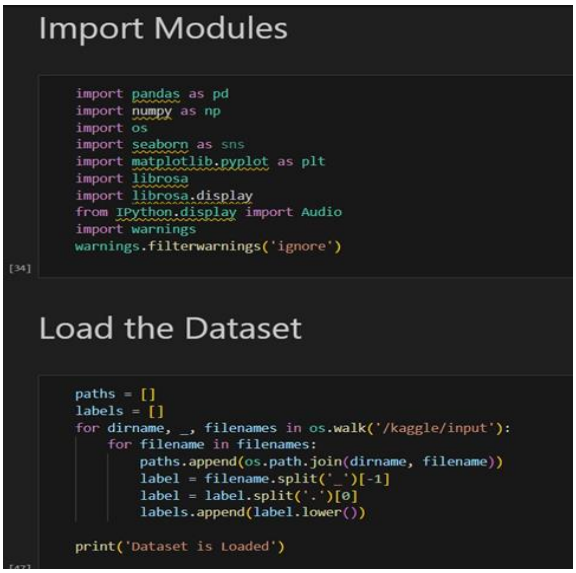

Figure 2: Importing Modules and Loading Dataset

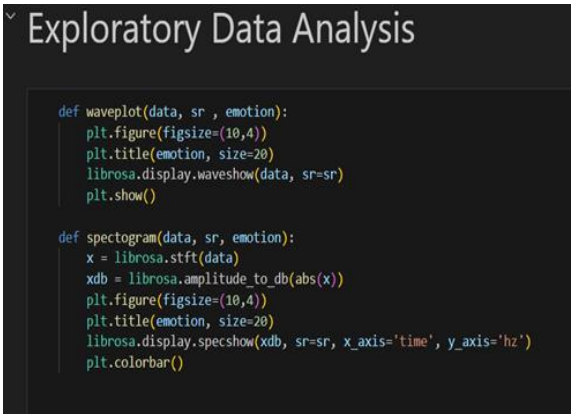

Figure 3: Exploratory Data Analysis

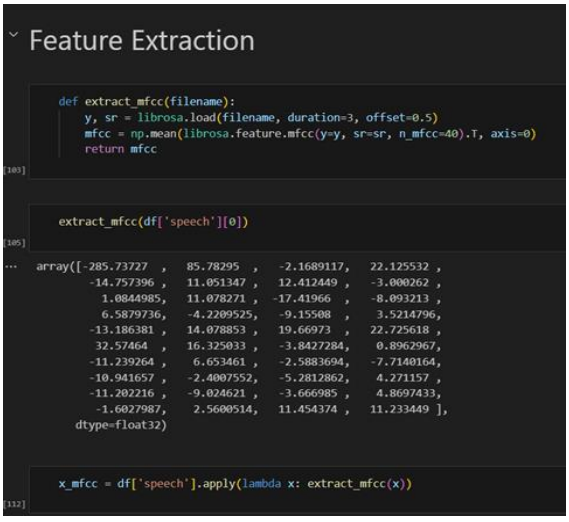

Figure 4: (a) Feature Extraction

Figure 4: (b) Feature Extraction

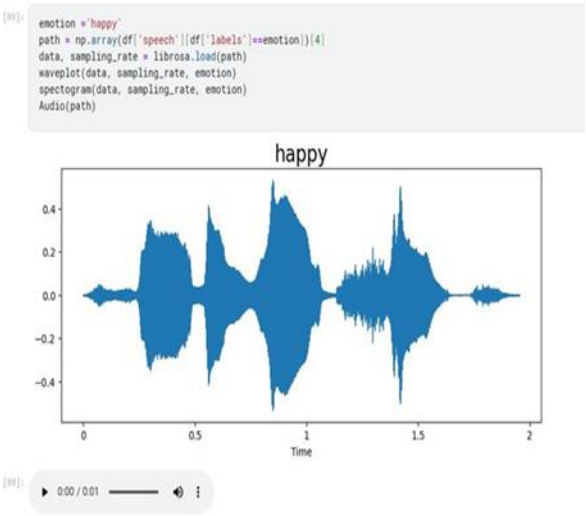

Figure 5:(a) Input with Happy Voice Note

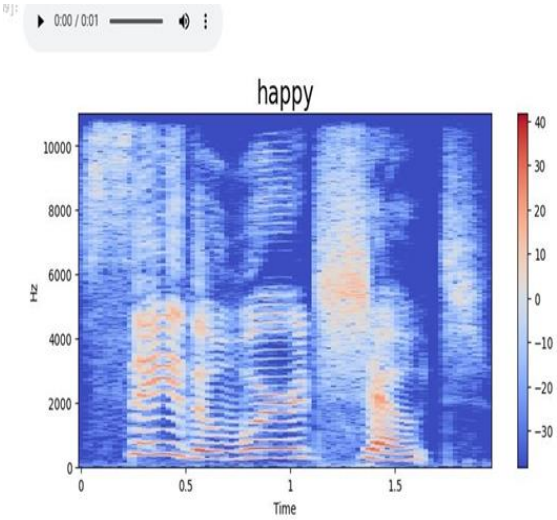

Figure 5: (b)Graphical Output of Happy Voice Note

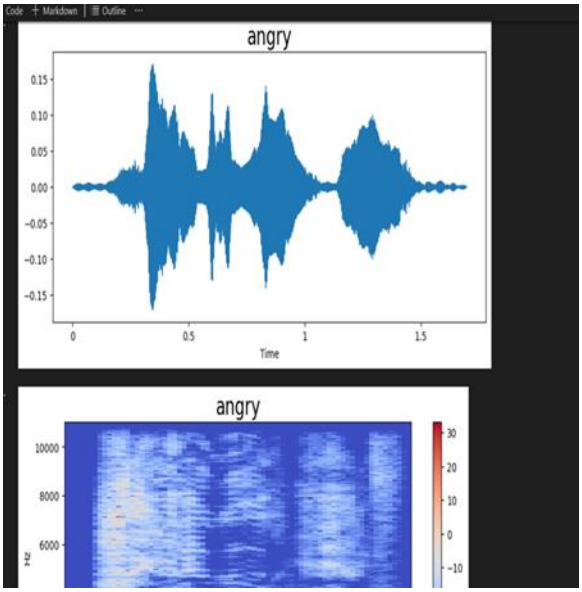

Figure 6: Graphical Input and Output of Angry Emotion

VI. DISCUSSION

The results of this study indicate that deep learning models, particularly LSTM, achieved higher accuracy compared to traditional machine learning approaches such as SVM. This improvement can be attributed to LSTM's ability to capture temporal dependencies in speech, which is crucial for identifying emotions effectively. The experimental findings align with existing research, demonstrating that deep learning-based models outperform traditional classifiers in Speech Emotion Recognition. Compared to prior studies, the use of MFCCs as a feature extraction method further enhanced classification accuracy by capturing essential frequency components of speech.

However, several challenges were encountered during the study. One major limitation was dataset imbalance, where certain emotions had significantly fewer samples than others, leading to biased predictions. Additionally, variations in speaker accents and background noise impacted model performance, highlighting the need for noise-robust feature extraction techniques. Another challenge was the computational complexity of deep learning models, which made real-time implementation more demanding. Despite these challenges, the study's findings demonstrate the effectiveness of deep learning in Speech Emotion Recognition and its potential for real-world applications. The ability to

accurately detect emotions from speech can be valuable in various domains, such as virtual assistants, healthcare, and affective computing, enabling more intelligent and emotionally aware human-computer interactions.

VII. CONCLUSION

This study successfully developed a Speech Emotion Recognition (SER) system using a dataset obtained from Kaggle. The system utilized various feature extraction techniques, including Mel-Frequency Cepstral Coefficients (MFCCs) and Spectrograms, to capture emotional cues from speech. Different classification models, including Support Vector Machines (SVM), Convolutional Neural Networks (CNN), and Long Short-Term Memory (LSTM) networks, were implemented and evaluated for their performance.

The results demonstrated that deep learning models, particularly LSTM, achieved the highest accuracy in recognizing emotions due to their ability to capture sequential dependencies in speech data. The study also highlighted the impact of dataset quality on model performance, emphasizing the importance of balanced and diverse data for effective emotion classification.

While the system performed well, challenges such as variations in speaker accents, background noise, and data imbalance were observed, affecting overall accuracy. These findings reinforce the significance of selecting appropriate feature extraction techniques and classification models to enhance emotion recognition from speech. The research contributes to the advancement of emotion-aware systems, improving their potential applications in human-computer interaction, virtual assistants, and affective computing

REFERENCES

[1] Ververidis, D., & Kotropoulos, C. (2006). "Emotional Speech Recognition: Resources, Features, and Methods". Speech Communication.

[2] Eyben, F., Wöllmer, M., & Schuller, B. (2010). "Opensmile – The Munich Versatile and Fast Open-Source Audio Feature Extractor". ACM Multimedia.

[3] Hassan, A., & Damper, R. I. (2013). "On Acoustic Emotion Recognition: Comparing Speech and Vocalizations". IEEE Transactions on Affective Computing

[4] Viswanath Ganapathy, Ranjeet K. Patro, Chandrasekhara Thejaswi, Manik Raina, Subhas K. Ghosh, Signal Separation using Time Frequency Representation, Honeywell Technology Solutions Laboratory

[5] Brani Vidakovic and Peter Mueller, Wavelets for Kids – A Tutorial Introduction, Duke University

[6] C. Valens, IEEE, A Really Friendly Guide to Wavelets, Vol.86, No. 11, Nov 2012

[7] Zhang, X., et al., "Deep Learning for Speech Emotion Recognition: A Review," IEEE Transactions on Affective Computing, 12(3), 2020,

[8] pp. 556–573.

[9] Schuller, B., et al., "Speech Emotion Recognition: Two Decades in a Nutshell, Benchmarks, and Ongoing Trends," Communications of the ACM, 61(5), 2018, pp. 90–103

[10] Jurafsky, D., and Martin, J. H., Speech and Language Processing, Pearson, 2019.

[11] Bishop, C. M., Pattern Recognition and Machine Learning, Springer, 2006.

[12] Understanding MFCCs: Features for Speech Recognition," Towards Data Science, Available at: https://towardsdatascience.com/understanding-mfcc s-features-for-speech-recognition-22d8bc2acbbc, Accessed on January 1, 2025.

[13] "Speech Emotion Recognition using Python," GeeksforGeeks, Available at: https://www.geeksforgeeks.org/speech-emotion-rec ognition-using-python/, Accessed on January 1, 2025.

[14] A. Hassan, R. Damper, and M. Niranjan, "On Acoustic Emotion Recognition: Compensating for Covariate Shift," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 7,

[15] pp. 1458-1468, 2013.

[16] P. Tzirakis, G. Trigeorgis, M. A. Nicolaou, B.

[17] W. Schuller, and S. Zafeiriou, "End-to-End Multimodal Emotion Recognition using Deep Neural Networks," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 8, pp. 1301-1309, 2017.

[18] E. Marchi, F. Ringeval, and B. Schuller, "Deep Recurrent Neural Networks for Audiovisual Emotion Recognition," in *Proceedings of the International Conference on Multimodal Interaction (ICMI)*, 2015, pp. 473-480.

[19] R. Cowie, E. Douglas-Cowie, and C. Cox, "Beyond Emotion Archetypes: Databases for Emotion Recognition," in *Proceedings of the International Conference on Affective Computing and Intelligent Interaction (ACII)*, 2005, pp. 909-914.

[20] C. Busso, M. Bulut, C. C. Lee, A. Kazemzadeh, and S. Narayanan, "IEMOCAP: Interactive Emotional Dyadic Motion Capture Database," *Journal of Language Resources and Evaluation*, vol. 42, no. 4, pp. 335-359, 2008.

[21] J. S. Kim, K. A. Lee, and I. H. Lee, "Emotion Recognition through Speech Using Multi-Layer Perceptron Models," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2018, pp. 5089-5093