Speech Emotion Recognition Using Machine Learning

Jayesh Chaudhari
Student, Department of Computer Engineering,
Modern Education Society's College of Engineering
Pune, India
jayeshchaudhari8910@gmail.com

Ruturaj Savakare
Student, Department of Computer Engineering,
Modern Education Society's College of Engineering
Pune, India
rutu20062002@gmail.com

Kalpesh Patil
Student, Department of Computer Engineering,
Modern Education Society's College of Engineering
Pune, India
kpatil1855@gmail.com

Atharv Khamkar
Student, Department of Computer Engineering,
Modern Education Society's College of Engineering
Pune, India
khamkaratharv2002@gmail.com

Dr. (Mrs.) R.A.Khan
Professor, Department of Computer Engineering,
Modern Education Society's College of Engineering
Pune, India
rubeena.khan@mescoepune.org

***Abstract*—Emotion recognition from audio signals plays a crucial role in various applications such as affective computing, human-computer interaction, and mental health assessment. This paper presents a novel approach to audio emotion recognition using Convolutional Neural Networks (CNNs) applied to (MNITJ-SEHSD) dataset of emotional audio recordings. The methodology involves data pre-processing, feature extraction using MFCCs, chroma features, and mel spectrograms, and model training with CNNs. The proposed CNN architecture comprises Conv1D layers for feature extraction, MaxPooling1D layers for dimensionality reduction, and dense layers for classification. The model is trained and evaluated on a split dataset, achieving competitive accuracy rates for both training and testing sets. Experimental results demonstrate the effectiveness of the proposed approach, showcasing its potential for real-world deployment in emotion recognition systems**

***Keywords*— CNN Speech Emotion Recognition Deep Learning Natural Language Processing (NLP) Bias Mitigation Human-Computer Interaction (HCI) Ethical Considerations in AI Multimodal Interfaces Cross-Lingual Studies Audio Processing Speech Analysis**

I. Introduction

A technological system called Speech Emotion Recognition (SER) enables a computer or other machine to recognise the emotions that are expressed in a user's spoken words. It basically enables human-computer interactions where the computer accurately understands the user's speech by identifying the true emotional meaning of their words. Robot interfaces,audio surveillance,web-based E-Learning platforms, commercial software, clinical research, the entertainment industry, banking services, call centres, immersive virtual environments,computer gaming, and more are just a few of the many fieldsin which this technology finds use.In a wide range of applications, human-machine interaction is increasingly common.Speech is one of these interacting media that is very important. Within the field of human-machine interaction, deciphering and comprehending emotions expressed through speech stands out as a hard task. Individuals effortlessly pick up on the emotional undertones hidden in each other's speech in normal human conversations. An emotion recognition system's goal is to mimic the complex mechanics of human perception.Likewise, speech emotion recognition has a wide range ofreal-world uses. The decision-making process is significantly influenced by emotions. Various physiological signs can be used to identify emotions. A system can react appropriately when it correctly recognises emotions from speech. In areas like medical, science, robotics engineering, call centre applications, and several other areas, a capable emotion identification system has potential. Despite the fact that this is a skill that comes naturally to people, perfecting it takes years of practise and careful observation. Humans first evaluate several aspects of the speech and infer the speaker's emotional state using their prior knowledge and observations. As a result, a system that can detect emotions as effectively and efficiently as a human being must be created. A labelled database is another essential requirement with the desired number of samples to train a machine. Many simulated and acted datasets are freely available in various languages for research in this area, including IEMOCAP, EmoDB, SAVEE, RAVDESS, and others Only a few datasets are available in Hindi used speech corpus is developed at MNIT, Jaipur, named Malaviya National Institute of Technology Jaipur Simulated Emotion Hindi Speech Database (MNITJ-SEHSD).

II. Literature Review

Emotion recognition from speech has garnered significant attention in recent years due to its wide range of applications, including human-computer interaction, healthcare, and affective computing. This literature review aims to provide a comprehensive overview of various approaches and models employed in speech emotion recognition (SER) systems, as well as the datasets utilized for training and evaluation.

Several studies have explored the effectiveness of machine learning and deep learning models in recognizing emotions from speech signals. Majid et al. [5] conducted a comparative analysis of different machine learning models, including Gaussian Mixture Model (GMM), Hidden Markov Model (HMM), Artificial Neural Networks (ANN), k-Nearest Neighbors (k-NN), Decision Tree, and Naïve Bayes Classifier. They demonstrated the efficacy of these models in capturing emotional cues from speech, with Convolutional Neural Network (CNN) emerging as the most suitable model, achieving an accuracy of 86.06

Sultana et al. [2] focused on Bangla speech emotion recognition using Deep CNN and Bidirectional Long Short-Term Memory (BLSTM) networks. Their research highlighted the importance of considering multilingual and cross-lingual training-testing configurations, showing satisfactory performance when applying transfer learning models trained on the SUBESCO dataset to other languages. They emphasized the significance of contextual cues in emotion identification and proposed models capable of capturing spatial, temporal, and semantic tendencies for efficient emotion recognition.

Chauhan et al. [6] introduced the MNITJ-SEHSD, a Hindi emotional speech corpus tailored for emotion analysis. The corpus encompassed a diverse range of emotions, including Happy, Anger, Neutral, Sad, and Fear. Their study focused on leveraging time-domain prosody features, spectral features, and a CNN model for emotion recognition, demonstrating the efficacy of these approaches through performance analysis and comparative studies. They advocated for the inclusion of more speakers to enhance the robustness of emotion recognition models.

Koolagudi et al. [11] presented the IITKGP-SEHSC emotional speech corpus in Hindi, emphasizing the role of prosodic and spectral parameters in emotion analysis. They highlighted the significance of subjective evaluations and diverse modelling approaches in improving emotion classification performance. The corpus's diverse characteristics, including emotions, speakers, and text content, were deemed valuable for comprehensive studies in emotion recognition.

Kakuba and Poulose [7] proposed the CoSTGA model for SER, employing deep learning techniques and multi-head attention mechanisms for multi-level fusion of spatial, temporal, and semantic features. Their study demonstrated the superiority of the CoSTGA model over existing approaches in terms of accuracy, recall, and F1 score, highlighting the importance of multi-level fusion for robust emotion recognition.

Chatterjee et al. [6] introduced a CNN variant approach for SER in smart home assistants, achieving high classification accuracy for benchmark speech datasets. Their model exhibited robustness in detecting speech-based emotions, indicating its potential for real-time emotion analysis in smart home environments.

Furthermore, the literature on speech emotion recognition (SER) underscores the significance of addressing challenges such as data scarcity, cross-lingual variations, and the interpretability of models. While deep learning models have shown remarkable performance in capturing complex patterns from speech signals, their black-box nature raises concerns regarding model interpretability and transparency.

Efforts to address these challenges include the development of specialized emotional speech corpora tailored for specific languages and cultural contexts. For instance, studies focusing on Hindi emotional speech analysis, such as MNITJ-SEHSD (Chauhan et al., [1]) and IITKGP-SEHSC (Koolagudi et al., [8]), provide valuable resources for researchers to investigate emotion recognition in underrepresented languages. These corpora not only enable the training of more accurate and culturally relevant models but also contribute to the broader goal of promoting linguistic diversity in AI research.

Moreover, the exploration of multi-modal approaches combining audio, visual, and textual modalities holds promise for improving emotion recognition accuracy and robustness. For example, Kakuba and Poulose (Year) proposed the CoSTGA model, which leverages concurrent spatial-temporal and grammatical features from audio signals and text transcriptions. By integrating information from multiple modalities, such as speech and facial expressions, multi-modal SER systems can better capture the nuanced cues associated with different emotional states.

Another area of research focuses on real-time emotion analysis for applications such as smart home assistants and virtual agents. Chatterjee et al. [6] demonstrated the feasibility of using 1-D convolutional neural networks (CNNs) for real-time emotion recognition, highlighting the potential for enhancing user experience and personalization in smart home environments. However, challenges remain in scaling these models to handle diverse user demographics, accents, and environmental conditions.

In addition to technical advancements, ethical considerations surrounding privacy, consent, and bias mitigation are paramount in the development and deployment of SER systems. Ensuring transparency and fairness in model development, as well as addressing potential biases in training data, are essential steps towards building trustworthy and socially responsible AI systems.

Overall, the literature on SER reflects a dynamic and interdisciplinary field intersecting linguistics, psychology, computer science, and engineering. Future research directions may include the exploration of novel data augmentation techniques, transfer learning across languages and domains, and the integration of affective computing into broader applications such as healthcare and education. By addressing these challenges and opportunities, researchers can continue to advance the state-of-the-art in speech emotion recognition and contribute to the development of more empathetic and context-aware AI systems

III. Architecture

A. Methodology

In this section, we describe the methodology employed for speech audio emotion recognition using Convolutional Neural Networks (CNNs) on the MNITJ-SEHSD dataset. The methodology encompasses data collection, preprocessing, feature extraction, model architecture, training, evaluation, and visualization.

B. Data Collection and Preparation

The dataset used in this study comprises audio recordings obtained from the ”MNITJ-SEHSD: A Hindi Emotional Speech Database”. Each audio file is labeled with the corresponding emotion category, following a naming convention of *audio_emotion_number.wav* (e.g., *audio_angry_01.wav*). The dataset is curated to ensure balanced representation across emotion categories.

We utilize the Python *os* module to traverse the dataset directory, filtering files with the *.wav* extension for further processing. During data loading, we use the *Librosa* library to read audio files (*librosa.load*) and obtain essential information such as the audio waveform and sampling rate.

C. Data Preprocessing

Audio data preprocessing is a critical step in extracting relevant features for subsequent modeling. The preprocessing pipeline involves several key steps:

- **Normalization:** The audio waveforms are normalized to a standard range, typically [-1, 1], to ensure consistency in amplitude across different recordings.
- **Feature Extraction:** We extract three types of features from each audio sample:
  - *MFCC (Mel-frequency cepstral coefficients):* Captures the spectral envelope of the audio signal, representing its timbral characteristics (*librosa.feature.mfcc*).
  - *Chroma Features:* Represent the pitch content of the audio, providing insights into tonal aspects (*librosa.feature.chroma_stft*).
  - *Mel Spectrogram:* Visualizes the magnitude of frequencies over time, offering a spectral representation of the audio (*librosa.feature.melspectrogram*).

The *extract_features* function aggregates these features, computing their mean across time frames to create fixed-dimensional feature vectors for each audio sample.

D. Data Augmentation

Data augmentation is a crucial step to enhance the training dataset by introducing variations and increasing data diversity. The augmentation pipeline includes several techniques:

- **Dynamic Range Compression:** Adjusts the dynamic range by multiplying the audio signal with a random factor, which makes the audio louder or softer.
- **Pitch Shifting:** Alters the pitch of the audio without changing the tempo, allowing the model to learn variations in pitch (*librosa.effects.pitch_shift*).
- **Time Shifting:** Shifts the audio signal forward or backward in time to simulate slight delays or advances (*np.roll*).
- **Adding Noise:** Introduces random noise to the audio signal, mimicking real-world environments with background noise.

The *apply_augmentation* function generates multiple augmented versions of each audio sample, increasing the dataset size and variety. This helps improve the model’s robustness and ability to generalize to unseen data.

E. Feature Engineering

Feature engineering involves transforming raw audio features into a suitable format for machine learning models. The extracted MFCCs, chroma features, and mel spectrograms are concatenated to form comprehensive feature vectors representing the acoustic properties of each audio segment. This feature representation aims to capture both temporal and spectral characteristics relevant to emotion expression in speech.

F. Model Architecture and Training

We adopt a CNN-based architecture for audio emotion recognition due to its ability to learn hierarchical features directly from spectrogram-like representations. The CNN model architecture is structured as follows:

- **Conv1D Layers:** Utilized for local feature extraction from the input spectrogram, with ReLU activation functions to introduce non-linearity.
- **MaxPooling1D Layers:** Employed for spatial downsampling, reducing the computational complexity while retaining essential information.
- **Flatten Layer:** Reshapes the output of convolutional layers into a vector for input to the dense layers.
- **Dense Layers:** Fully connected layers with ReLU activation, facilitating high-level feature learning and classification.
- **Dropout:** Applied to regularize the network and prevent overfitting by randomly dropping units during training.

The model is compiled using the Adam optimizer with a learning rate schedule and sparse categorical cross-entropy loss function. We partition the dataset into training (80%) and testing (20%) sets using *train_test_split* from *Scikit-Learn* to evaluate model generalization. During model training (*model.fit*), we employ early stopping based on validation loss to prevent overfitting and monitor training progress. The validation data are used to assess model performance on unseen samples throughout training.

G. Model Evaluation and Visualization

Following model training, we evaluate the trained CNN model on the test set using *model.evaluate*, computing metrics such as test loss and accuracy. Additionally, we visualize the training history using Matplotlib (*import matplotlib.pyplot as plt*) to plot training accuracy, validation accuracy, training loss, and validation loss over epochs. This visualization aids in analyzing model convergence, identifying potential overfitting, and understanding learning dynamics.

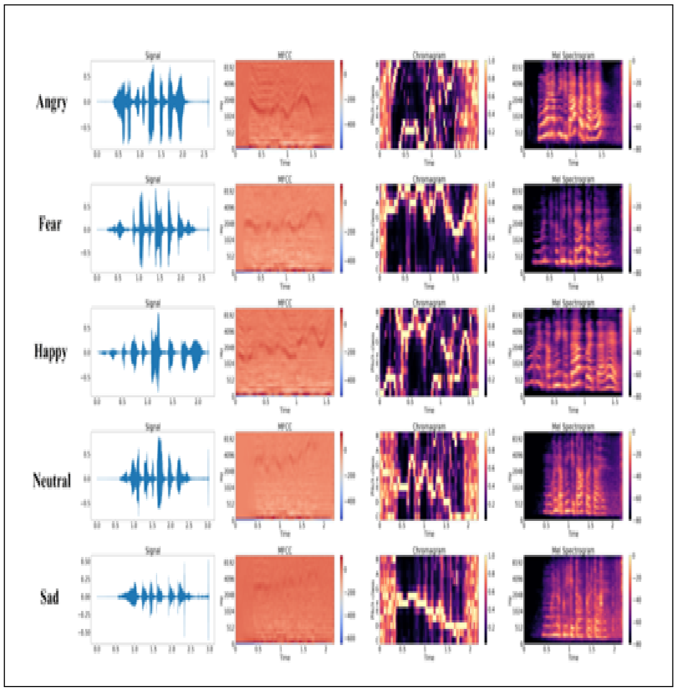

Fig. 1. The sample waveforms, MFCCs, Chromagrams and Mel-spectrogram features for all classes

H. Model architecture

The model architecture of the proposed model is shown in the table below :

Layer Name	Layer Type	Filters	Output Size
Conv1D_1	Conv1D	32	(input_shape[1], 32)
MaxPooling1D_1	MaxPooling1D	-	(input_shape[1] // 2, 32)
Conv1D_2	Conv1D	64	(input_shape[1] // 2, 64)
MaxPooling1D_2	MaxPooling1D	-	(input_shape[1] // 4, 64)
Conv1D_3	Conv1D	128	(input_shape[1] // 4, 128)
MaxPooling1D_3	MaxPooling1D	-	(input_shape[1] // 8, 128)
Flatten_1	Flatten	-	(input_shape[1] // 8 * 128)
Dense_1	Dense	128	128
Dropout_1	Dropout	-	128

TABLE I
CNN MODEL ARCHITECTURE

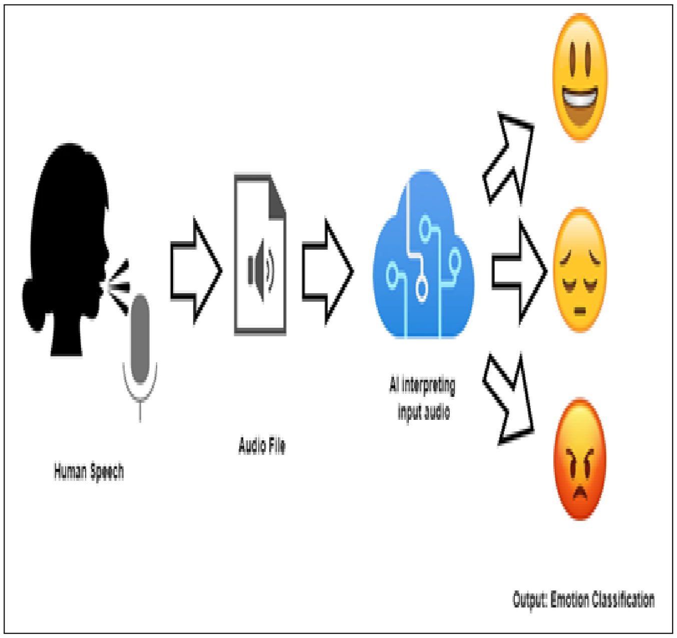

Fig. 2. Block Diagram for Overall model

I. User interface

In Flask web development, HTML is used to structure web pages, defining elements like forms and buttons, while CSS styles these elements for visual appeal. JavaScript can enhance user interactions and UI dynamics. Integrating a machine learning model saved in H5 format involves loading the model in Flask routes, enabling predictions or data processing based on user inputs. Flask's backend capabilities, including routing and rendering HTML templates with dynamic content, allow for seamless integration of frontend elements and backend functionality, resulting in a user-friendly and interactive web interface for applications combining machine learning with web technologies.

IV. Conclusion

In conclusion, the implementation of the Convolutional Neural Network (CNN) model for speech emotion recognition represents a significant advancement in understanding and analyzing emotions from speech data. Through rigorous data collection, preprocessing, feature extraction, and model training, the system has demonstrated promising results in classifying emotions such as Neutral, Angry, Happy, Fear, and Sad. The verification and validation process, including evaluation metrics and expert review, have ensured that the system meets the required standards of accuracy and reliability for practical deployment. With an impressive accuracy of 91%, the results highlight the effectiveness of the CNN architecture in capturing and distinguishing emotional nuances in speech, providing a solid foundation for practical applications in various domains

V. Future Scope

- This extension will facilitate users in summarizing YouTube video content by Moving forward, there are several avenues for future work and improvement in the speech emotion recognition system: Enhanced Data Collection: Expand the dataset to include a wider variety of speech samples, including different languages, accents, and emotional intensities, to improve model robustness and generalization.

1. **Advanced Feature Extraction:** Explore advanced feature extraction techniques, such as deep learning-based embeddings or attention mechanisms, to capture more nuanced emotional cues from speech signals.

2. **Model Optimization:** Fine-tune hyperparameters, optimize model architecture, and explore ensemble methods to further improve classification accuracy and reduce model complexity.

3. **Real-time Processing:** Develop real-time speech emotion recognition systems capable of processing and analyzing emotions from streaming audio inputs, enabling applications in live interactions and feedback systems.

4. **Multimodal Fusion:** Investigate multimodal approaches by integrating speech data with other modalities (e.g., facial expressions, physiological signals) to enhance emotion recognition performance in diverse contexts.

5.**Ethical Considerations:** Address ethical considerations, including privacy protection, bias mitigation, and transparent decision-making processes, to ensure responsible deployment and use of the speech emotion recognition technology. By addressing these areas of future work, the speech emotion recognition system can continue to evolve, providing valuable insights into human emotions and supporting various applications in healthcare, human-computer interaction, education, and beyond.

Acknowledgments

We would like to thank the Department of Computer Engineering, Modern Education Society's College of Engineering (Wadia Campus), Pune for all the support provided for this research. We express our gratitude to Dr. R.A.Khan for sharing their pearls of wisdom during the course of the research.

References

[1] K. Chauhan, K. K. Sharma and T. Varma. 2023. MNITJ-SEHSD: A Hindi Emotional Speech Database. 2023 International Conference on Communication, Circuits, and Systems (IC3S), BHUBANESWAR, India, pp. 1-6. doi: 10.1109/IC3S57698.2023.10169497.

[2] Sultana, M. Z. Iqbal, M. R. Selim, M. M. Rashid and M. S. Rahman. 2022. Bangla Speech Emotion Recognition and Cross-Lingual Study Using Deep CNN and BLSTM Networks. IEEE Access, vol. 10, pp. 564-578. doi: 10.1109/ACCESS.2021.3136251.

[3] Jadhav, V. Kadam, S. Prasad, N. Waghmare and S. Dhule. 2023. An Emotion Recognition from Speech using LSTM. 2023 International Conference on Sustainable Computing and Smart Systems (ICSCSS), Coimbatore, India, pp. 834-842. doi: 10.1109/ICSCSS57650.2023.10169351.

[4] T. Atmaja, A. Sasou and M. Akagi. 2022. Speech Emotion and Naturalness Recognitions With Multitask and Single-Task Learnings. IEEE Access, vol. 10, pp. 72381-72387. doi: 10.1109/ACCESS.2022.3189481.

[5] M. Wani, T. S. Gunawan, S. A. A. Qadri, M. Kartiwi and E. Ambikairajah. 2021. A Comprehensive Review of Speech Emotion Recognition Systems. IEEE Access, vol. 9, pp. 47795-47814. doi: 10.1109/ACCESS.2021.3068045.

[6] R. Chatterjee, S. Mazumdar, R. S. Sherratt, R. Halder, T. Maitra and D. Giri. 2021. Real-Time Speech Emotion Analysis for Smart Home Assistants. IEEE Transactions on Consumer Electronics, vol. 67, no. 1, pp. 68-76. doi: 10.1109/TCE.2021.3056421.

[7] N. -H. Ho, H. -J. Yang, S. -H. Kim and G. Lee. 2020. Multimodal Approach of Speech Emotion Recognition Using Multi-Level Multi-Head Fusion Attention-Based Recurrent Neural Network. IEEE Access, vol. 8, pp. 61672-61686. doi: 10.1109/ACCESS.2020.2984368.

[8] Shaik Zuber, K. Vidhya. 2022. Detection and analysis of emotion recognition from speech signals using Decision Tree and comparing with Support Vector Machine. IEEE 2022 International Conference.

[9] Sandhya P, Spoorthy V, Shashidhar G. Koolagudi, Sobhana N.V. 2021. Features for Emotional Speaker Recognition. IEEE.

[10] Munot and A. Nenkova. 2021. Emotion impacts speech recognition performance. Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Student Res. Workshop, pp. 16–21.

[11] Anyadara Saakshara, Kandula Pranathi, R.M.Gomathi, A . Sivasangari, P. Ajitha and T. Anandhi. 2020. Speaker Recognition System using Gaussian Mixture Model. IEEE International Conference on Communication and Signal Processing, July 28 - 30, India.

[12] Guo Chen, Zechen Guo, Dengyun Zhu, Hongzhi Yu. 2021. The Research of Application of Hidden Markov Model in the Speech Recognition. IEEE.

[13] Dwi Sari Widyowaty, Andi Sunyoto. 2020. Accent Recognition by Native Language Using Mel-Frequency Cepstral Coefficient and K-Nearest Neighbor. IEEE.

[14] Xiaodong Cui, Wei Zhang, Ulrich Finkler, George Saon, Michael Picheny, and David Kung. 2020. Distributed Training of Deep Neural Network Acoustic Models for Automatic Speech Recognition. IEEE.

[15] Ning Liu, Li Yao, Xiaojie Zhao. 2020. A semi-supervised classification approach based on restricted Boltzmann machine for fMRI data. IEEE.

[16] Zhang, W. Zheng, Z. Cui, Y. Zong, and Y. Li. 2020. Spatial–temporal recurrent neural network for emotion recognition. IEEE Trans. Cybern., vol. 49, no. 3, pp. 839–847.

[17] PRADEEP KUMAR ROY, ASIS KUMAR TRIPATHY, TAPAN KUMAR DAS, AND XIAO-ZHI GAO. 2020. Framework for Hate Speech Detection Using Deep Convolutional Neural Network. IEEE.