International Journal of Multidisciplinary Research and Growth Evaluation.

Speech Emotion Recognition

Kondaparthy Vaishnavi [1], Myana Vaishnavi [2*], Dr K. Vaidehi [3]
[1-2] BE, AI&DS, VIII Sem., SCETW, Hyderabad, India
[3] Professor & Head Department of ADCE, SCTEW, OU, Hyderabad, India

* Corresponding Author: **Myana Vaishnavi**

Article Info

ISSN (online): 2582-7138
Volume: 06
Issue: 02
March-April 2025
Received: 18-02-2025
Accepted: 23-03-2025
Page No: 1173-1192

Abstract
The goal of speech emotion recognition (SER) is to recognise and categorise emotional states expressed by speech signals, improving applications in healthcare, education, customer service, and human-computer interaction. Recent developments in SER are the main topic of this review, with a special emphasis on deep learning techniques that combine verbal and auditory data. Despite the moderate success of traditional methods based on prosody and hand-crafted features like Mel Frequency Cepstral Coefficients (MFCC), deep learning models like Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) networks greatly enhance the capacity to capture intricate, emotion-specific speech patterns. Well-known datasets like as IEMOCAP and RAVDESS are frequently used to assess the performance of SER systems, and accuracy metrics are employed to compare the efficacy of different strategies.
Accurately identifying spontaneous emotional speech, which lacks the structured cues seen in performed emotions, and maintaining robustness in a variety of acoustic settings are two of the main hurdles in SER. To solve these problems, recent research has incorporated sophisticated neural architectures, including hybrid CNN-LSTM models and attention mechanisms. Additionally, self-supervised learning techniques have become popular choices for scenarios involving little labelled data and low-resource languages. These models increase the accuracy of emotion categorisation and improve generalisation across languages and speakers by utilising both labelled and unlabelled data.
Another possible direction for future SER research is the integration of multimodal data, such as merging audio with textual or visual information. More complex emotion recognition may be possible with models that use context-aware techniques like sentiment-weighted attention. It is anticipated that SER technology will make a substantial contribution to emotionally aware AI systems as it develops, offering responsive, adaptive interactions that have a deeper comprehension of human emotions.

Keywords: Speech Emotion Recognition, Deep Learning, Convolutional Neural Networks, Long Short-Term Memory, Self-Supervised Learning, Attention Mechanisms, Multimodal Data, Acoustic Features

1. Introduction

Speech Emotion Recognition is a revolutionary combination of the complex field of speech processing with cutting-edge technologies. This new technique uses computing power to change the way emotions are identified in speech, making it faster, more accurate, and more effective. By evaluating big voice datasets, identifying emotional patterns, and refining identification models, voice Emotion identification holds the potential to completely transform human-computer interaction. Speech Emotion Recognition helps machines better perceive and react to human emotions in a more intuitive and natural way by bridging the gap between human emotions and technology.
In a world where technology and human connection are changing together, machines' capacity to comprehend human emotions is a ground-breaking development. In order to extract emotions from speech, the cutting-edge interdisciplinary discipline of speech emotion recognition (SER) integrates psychology, speech processing, and artificial intelligence.

In contrast to conventional speech recognition, which only considers linguistic content, SER seeks to understand a speaker's underlying emotional state in order to improve the effectiveness and intuitiveness of human-machine communication. With applications ranging from smart assistants and security systems to healthcare and customer service, SER is transforming human-machine interaction and giving artificial intelligence a new dimension.

The foundation of speech emotion recognition is the notion that several aspects of speech, including pitch, tone, intensity, and rhythm, represent human emotions. Machines are able to recognise emotions such as joy, sorrow, rage, fear, and surprise by examining these characteristics. AI-driven systems become more sympathetic and user-responsive when they can identify emotions in speech, improving human-computer interaction.

Speech signal acquisition, feature extraction, emotion categorisation, and output interpretation are the usual steps in a SER system. The audio input is initially captured, and then pertinent properties including pitch, energy, and Mel-Frequency Cepstral Coefficients (MFCCs) are extracted. The retrieved features are subsequently categorised into several emotion categories by machine learning or deep learning algorithms, offering valuable insights into the speaker's emotional state.

SER uses a variety of methods, from deep learning models to traditional machine learning techniques. While recent deep learning architectures use Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and transformers to improve recognition accuracy, traditional techniques include Support Vector Machines (SVM), Hidden Markov Models (HMM), and Decision Trees. These methods enhance emotion detection skills by utilising large datasets and processing capacity.

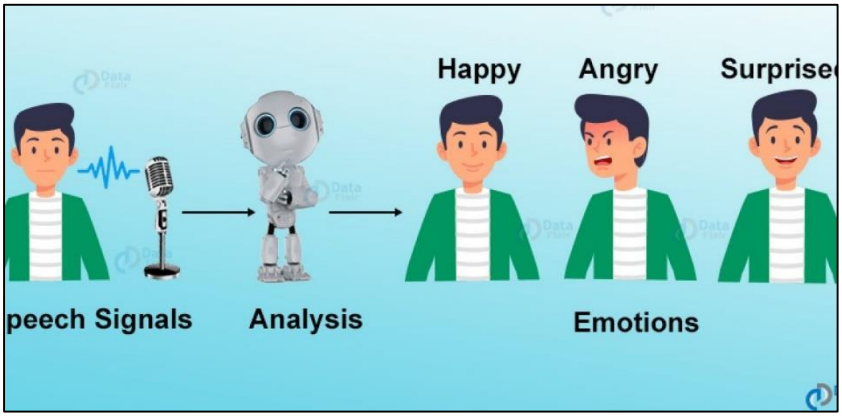

Fig 1: Speech Emotion Recognition

Even with its progress, SER still has a number of obstacles to overcome. Emotion identification is complicated by speech variability brought on by various languages, dialects, and speaking styles. Accuracy is also impacted by outside environmental influences, background noise, and overlapping speech. Furthermore, it is challenging to develop a universal emotion detection model because emotions are subjective and might manifest themselves differently in various people.

Applications for speech emotion recognition are numerous and span many industries. It helps in the diagnosis of mental health conditions like anxiety and depression in the medical field. SER improves chatbot interactions in customer care, guaranteeing a better user experience. Additionally, it is employed in entertainment to improve interactive gaming experiences and in security and forensics to identify stress or dishonesty in speech.

Deep learning developments and the use of multimodal emotion identification bode well for SER's future. The accuracy of emotion identification can be greatly increased by combining voice data with physiological signs, facial expressions, and environmental information. Furthermore, personal assistant technologies are about to undergo a revolution thanks to the introduction of real-time SER in wearable technology and smartphone applications.

Speech Emotion Recognition is a game-changing technique that connects artificial intelligence and human emotions. SER improves how machines comprehend and react to human emotions by utilising cutting-edge speech processing methods and machine learning models. Even if there are still obstacles to overcome, ongoing study and development in this area bode well for a time when machines will be able to communicate with people more naturally and sympathetically, leading to previously unheard-of technological breakthroughs.

The goal of Speech Emotion Recognition (SER) technology is to detect emotions in speech by using voice traits such as pitch, tone, intensity, and rhythm to deduce the emotional state of the speaker. Accurately detecting emotions is the system's goal; this is a skill that is becoming more and more useful in fields like customer service, healthcare, and human-computer interaction. The development of SER has been hastened by recent developments in machine learning and signal processing, allowing for more precise, context-sensitive emotion detection that closes the empathy gap between humans and machines.

1.1. History of Speech Emotion Recognition:

Speech Emotion Recognition's history begins with the early research on human emotions and speech patterns conducted in the middle of the 20th century. The groundwork for computational methods was laid by linguistics and psychology researchers who started examining the connection between emotional states and vocal expressions. Due to a lack of computing power, early research mostly depended on manual speech feature analysis. But as signal processing capabilities improved, the concept of automating speech emotion recognition grew in popularity.

The emergence of digital signal processing and machine learning by the late 20th century resulted in the creation of

the first computational models for speech emotion recognition. To categorise emotions, researchers started employing feature extraction methods like pitch, formants, and energy. Early systems used traditional machine learning models, such as Support Vector Machines (SVM), Gaussian Mixture Models (GMM), and Hidden Markov Models (HMM), which had a moderate level of effectiveness in controlled settings. These techniques showed that it is possible to use algorithms to identify emotions in voice signals, despite their drawbacks.

Speech Emotion Recognition changed dramatically in the twenty-first century with the introduction of deep learning. By recognising intricate speech patterns, neural networks—in particular, Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs)—made it possible to detect emotions with greater accuracy. SER systems' performance was further improved by large datasets and increased processing power, which increased their applicability to real-world situations. In order to increase recognition accuracy, researchers also investigated multimodal techniques, which combine speech with physiological signs and facial expressions.

Research on Speech Emotion Recognition is now underway, with applications in a number of sectors, such as human-computer interaction, healthcare, customer service, and security. The possibilities of SER are constantly being pushed further by developments in artificial intelligence, especially transformer models and self-supervised learning. A future where machines can comprehend and react to human emotions with astonishing precision is promised by ongoing research and technology developments, despite obstacles like data unreliability, cultural differences, and ethical concerns.

1.2. Overview

Speech Emotion Recognition is a device that can read and figure out how people are feeling from the way they talk. It looks at different parts of speech, like pitch, tone, energy, and rhythm, to figure out how someone is feeling. It improves virtual helpers, customer service, mental health assessments, and human-computer connection by closing the gap between how people talk to computers and how computers understand them. Speech Emotion Recognition keeps changing as technology gets better, making it more accurate and reliable in a wide range of real-life situations.

Emotion recognition is much more accurate and quick now that Speech Emotion Recognition with deep learning is used. Deep learning models like Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and Long Short-Term Memory (LSTM) networks make it easier to record complex speech data, which helps machines better understand how people are feeling. These models look at a lot of data and learn complicated patterns in tone, pitch, and strength that are important for figuring out how someone is feeling. Deep learning is different from other types of machine learning because it doesn't require human feature extraction. This means that raw audio data can be used for training from start to finish. Speech Emotion Recognition is also being pushed to its limits by transformer models and self-supervised learning methods.

1.3. Aim, Scope and Objectives:

Aim: The goal of this project is to use deep learning models to make a Speech Emotion Recognition system that can correctly classify feelings from speech data. Using advanced neural network architectures like Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), the system will look at speech features to accurately detect emotional states. This will make uses in healthcare, human-computer interaction, and sentiment analysis better.

Scope: This project's goal is to create and use deep learning methods to create a Speech Emotion Recognition system. It includes cleaning up speech data, pulling out the important parts, teaching deep learning models on big sets of emotional speech, and testing how well they do. The project will focus on real-time apps, putting SER to use in virtual helpers, robots that can understand emotions, and systems that keep an eye on mental health.

Objectives: The main goal of this project is to build and implement a deep learning-based Speech Emotion Recognition model that successfully classifies different emotional states from speech. Specific objectives include collecting and preprocessing a large dataset of emotional speech recordings, extracting key speech features, implementing CNNs and RNNs for emotion classification, optimizing model performance, and evaluating the system's accuracy in real-world scenarios. Additionally, the project aims to explore multimodal approaches by integrating speech data with facial expressions or physiological signals to enhance recognition accuracy.

1.4. Importance of SER in Technology

SER has significant uses in many different fields of technology. By enabling social robots and virtual assistants to react adaptively depending on a user's emotional state, emotion-aware systems improve user experiences. This is particularly helpful in customer service. Additionally, SER can be very helpful in mental health monitoring, helping to identify changes in emotions that may be signs of illnesses like anxiety or depression. SER can be used in education and adaptive learning to measure emotional reactions and student engagement, allowing for tailored support and feedback.

1.5. Challenges in Speech Emotion Recognition

The field of speech emotion recognition (SER) is also challenged by the intrinsic subjectivity of emotions. Different individuals may express the same emotion in varied ways, influenced by personal experiences, cultural norms, and linguistic differences. For example, anger in one culture may be expressed with loud, sharp tones, while in another, it may be conveyed with a calmer yet firm voice. This variability makes it difficult to develop universal models that can accurately detect emotions across diverse populations. As a result, researchers are exploring ways to incorporate cross-cultural datasets and domain adaptation techniques to improve the generalizability of SER systems.

Another obstacle is the dynamic nature of emotions, which are not always static and can evolve over the course of a conversation. A speaker may transition from frustration to relief within a short span, making it challenging for SER models to capture these emotional shifts effectively. Traditional models often rely on isolated audio segments, missing the broader context of the dialogue. To address this, newer approaches integrate temporal modeling techniques, such as attention mechanisms and long short-term memory (LSTM) networks, which allow models to track emotional progressions more accurately. By considering context and temporal dependencies, SER models can provide a more holistic understanding of emotions.

Furthermore, advancements in multimodal emotion recognition are enhancing SER capabilities by combining audio with other data sources like facial expressions and physiological signals. Since emotions are often expressed through multiple channels, integrating information from speech, facial cues, and even biometric data like heart rate or skin conductance can significantly boost accuracy. Deep learning techniques, including transformer-based models and self-supervised learning, are being employed to fuse these modalities effectively. As research progresses, the future of SER lies in developing robust, context-aware models that can perform well in real-world applications, from mental health monitoring to human-computer interaction**.**

2. Literature Survey

[1] In order to enhance automated speech recognition (ASR) for child speech, Jain *et al*. investigated the use of wav2vec2 in conjunction with self-supervised learning (SSL). SSL enables pretraining on unlabelled adult data and fine-tuning with little child data, which is advantageous for ASR models that usually require large amounts of labelled data. This method demonstrated how well SSL adapts models for distinct speech characteristics by drastically lowering the Word Error Rate (WER) on child-specific datasets. Other specialised speech applications, such as emotion recognition across age groups, may benefit from this technique.

[2] Sgouros *et al*. introduced a new music source separation (MSS) system that combines an attentive multiresUNet with the Short-Time Discrete Cosine Transform (STDCT). By using real-valued data, this technique solves the phase recovery problem and effectively separates musical elements like bass and vocals. The system demonstrated potential for real-time audio editing and production applications when tested on the MUSDB18 dataset, yielding results comparable to state-of-the-art models with less computing complexity.

[3] To improve emotion recognition from speech, Santoso *et al*. created a self-attention weight adjustment technique that takes into account both textual and auditory characteristics. The approach prioritises more dependable acoustic input where ASR faults could degrade text quality by integrating confidence measures into the self-attention process. Experiments on the IEMOCAP dataset demonstrated that this method performed better than earlier models and offered a reliable solution for emotion recognition, particularly in real-time processing applications.

[4] Because each speech impediment is unique, Yu *et al*. suggested the MAV-HuBERT audio-visual fusion architecture to enhance ASR for dysarthric speech. The model adjusts pre-trained HuBERT models to dysarthric speakers by combining audio data with facial speech aspects. The system's considerable improvement in WER on the UASpeech dataset demonstrated the promise of multi-modal ASR models for people with speech impairments.

[5] Using a deep learning methodology and a feature set comprising MFCC, Zero Crossing Rate, and Harmonic to Noise Ratio, Aouani and Ben Ayed examined emotion identification. They showed that on the RML dataset, integrating these features with Support Vector Machines (SVM) produces great accuracy. Their research highlights how deep learning might enhance emotion recognition in domains where adaptive reactions to emotional states are useful, such as healthcare, education, and automobile technology.

[6] In order to increase SER accuracy, Kumbhar and Bhandari used Mel-Frequency Cepstral Coefficients (MFCC) in conjunction with an LSTM network, taking use of LSTM's capacity to recognise temporal patterns. The model's 84.8% accuracy on the RAVDESS dataset demonstrated that LSTM's temporal dynamics make it a good choice for speech emotion recognition, particularly in applications involving real-time human-computer interaction.

[7] For improved emotion recognition, Wang *et al*. presented a dual-sequence LSTM model that independently processes MFCC and mel-spectrogram information. Their method improves accuracy on datasets such as IEMOCAP by utilising complementing information from both feature types. The architecture of this model shows how effective multi-stream processing is in extracting subtle emotional information from speech.

[8] For low-power ASR applications on FPGA platforms, Yin *et al*. created a spiking LSTM accelerator that lowers energy consumption without sacrificing processing speed. The model's creative design reduces power consumption without sacrificing performance by substituting simpler operations for conventional multiplication. This concept is particularly applicable to wearable technology and mobile ASR systems, which are used in energy-sensitive devices.

[9] In order to improve emotion recognition, Kadiri *et al*. concentrated on excitation characteristics, such as basic frequency, excitation strength, and excitation energy. They showed that excitation traits offer a unique method of classifying emotions, producing good results across languages, by contrasting emotional speech with neutral references. By catching characteristics that conventional spectral-based approaches frequently miss, this method expands the application of SER.

[10] A multiscale deep convolutional LSTM model was presented by Zhang *et al*. to identify emotions in spontaneous speech. Their method captures a variety of emotional cues by processing mel-spectrograms of different lengths, and it achieves competitive performance on the AFEW5.0 and BAUM-1s datasets. Because of its versatility, this model can be used in real-world scenarios where improving machine comprehension of human affect requires genuine emotional displays.

[11] For Speech Emotion Recognition (SER), Abdelwahab and Busso investigated the use of active learning (AL) to enhance model performance with sparse labelled data. Their method efficiently maximises feature variety and minimises labelling effort by using uncertainty-based sampling to choose the most informative training examples. Tests conducted on the MSP-Podcast dataset demonstrated that AL greatly improves cross-domain generalisation. This study demonstrates how AL can improve the effectiveness and adaptability of SER models in settings with little emotional data.

[12] The capacity of the model to learn features directly from log-spectrograms was the main focus of Zheng *et al*.'s investigation into the use of Deep Convolutional Neural Networks (DCNNs) for emotion identification. The DCNN architecture achieves a significant gain in accuracy over SVM-based approaches on the IEMOCAP dataset by extracting high-level emotional representations from spectral data, in contrast to standard hand-crafted features. The efficiency of deep learning in SER is demonstrated by this work, especially for jobs that call for reliable feature extraction over a range of emotional expressions.

[13] In their comparative analysis of different machine learning models for SER, Kerkeni *et al*. concentrated on

feature extraction methods such as Modulation Spectral Features (MSF) and Mel-Frequency Cepstral Coefficients (MFCC). RNNs performed the best on the Spanish dataset, according to their testing of models like SVMs and RNNs on the Berlin and Spanish emotional databases. The study emphasises how crucial it is to select features and models that are suitable for each dataset in order to achieve successful SER.

[14] In their unique SER framework, Mustaqeem *et al*. integrate CNN for feature extraction, BiLSTM for temporal learning, and clustering-based key segment selection. On datasets such as IEMOCAP and EMO-DB, the method improves classification accuracy and computing efficiency by processing emotional parts selectively. By optimising both accuracy and processing burden, this method shows how useful it is to selectively focus on important portions for emotion recognition.

[15] Bypassing conventional preprocessing and putting raw audio straight into the network, Qayyum *et al*. investigated the use of CNNs in SER. On the SAVEE dataset, their CNN architecture demonstrated great accuracy, demonstrating CNNs' capacity to extract high-level emotional signals from unprocessed audio without the need for intensive feature engineering. For real-time applications, this approach is beneficial, especially in settings where simplicity and speed are crucial.

[16] Mesaros *et al*. examined the outcomes of the DCASE 2016 challenge, which concentrated on challenges related to Sound Event Detection (SED) and Acoustic Scene Classification (ASC). Given that CNN and RNN models performed better on standardised datasets, the challenge demonstrated the trend towards deep learning. This work establishes a standard for SER applications in context-aware systems going forward by highlighting the effectiveness of deep learning in auditory scene analysis.

[17] With an emphasis on noise resilience, Satt *et al*. created a CNN-LSTM model for effective emotion recognition from spectrograms. On the IEMOCAP dataset, their model detected emotions like melancholy and rage with 68% accuracy after removing non-harmonic components. The model is appropriate for real-world applications in human-computer interaction, where environmental influences frequently compromise the accuracy of emotion recognition, due to its robustness to noise.

[18] An end-to-end deep neural network model that directly processes unprocessed audio inputs was presented by Tzirakis *et al*. for continuous emotion identification. The model captures subtle temporal variations in emotions by combining CNN for feature extraction with LSTM layers for temporal dynamics. It performed better than conventional techniques when tested on the RECOLA dataset, demonstrating the benefit of continuous emotion prediction for applications needing real-time responsiveness.

[19] To increase classification accuracy, Sun developed an end-to-end SER system with gender-specific inputs. The model improved accuracy by 7% on Mandarin, German, and English datasets when raw audio was processed with a gender-adaptive layer using residual CNNs. This method highlights the significance of demographic factors in improving SER performance while successfully addressing gender-related vocal disparities.

[20] To capture both temporal and spatial emotional cues in mel-spectrograms, Zhao *et al*. integrated Fully Convolutional Networks (FCN) and Bidirectional LSTM (BLSTM) with attention mechanisms. Their model, which benefited from attention mechanisms that highlight emotionally salient elements, obtained considerable accuracy increases on IEMOCAP. This method demonstrates the effectiveness of deep spectrum representations for precise emotion identification in a variety of datasets.

[21] FastSpeech, a non-autoregressive text-to-speech (TTS) model developed by Ren *et al*., allows for precise and quick voice synthesis. FastSpeech uses parallel processing, which drastically cuts down on inference time compared to sequential TTS models. Its architecture, which is based on the Transformer model with a length regulator, directly controls prosody and speed by aligning phoneme duration. This invention addresses common TTS problems like repeated or skipped words by enabling precise modifications to voice qualities. FastSpeech is appropriate for real-time applications needing quick, reliable, and controlled TTS since it reached quality comparable to cutting-edge autoregressive models.

[22] Convolutional Neural Networks (CNNs) were used by Mao *et al*. to extract emotion-relevant features from speech data, with an emphasis on learning representations that are resistant to background noise and speaker fluctuation. The model captures hierarchical feature representations through salient discriminative feature analysis and a sparse autoencoder. Their strategy achieved great accuracy across many datasets, outperforming conventional feature extraction techniques. This work demonstrates CNN's ability to improve SER, even under difficult circumstances, giving it a practical solution for emotion identification in the real world.

[23] To overcome the drawbacks of unimodal methods, Yoon, Byun, and Jung devised a multimodal SER system that incorporates both textual and auditory input. While the text encoder retrieves semantic information from transcripts, the model uses gated recurrent unit (GRU) networks and Mel-frequency cepstral coefficients (MFCCs) to record audio features. The final dual encoder demonstrates that multimodal input allows for a better understanding of emotional expressions in speech, which is advantageous for sentiment-sensitive applications. It achieves state-of-the-art accuracy on the IEMOCAP dataset, especially for complicated emotions.

[24] To enhance SER on constrained datasets, Abdelhamid *et al*. introduced a hybrid CNN-LSTM model optimised using a Stochastic Fractal Search-guided Whale Optimisation Algorithm (SFS-WOA). While LSTM analyses temporal patterns, CNN layers record spatial data, and SFS-WOA adjusts the model's hyperparameters. This method proved resilient to overfitting problems and showed good classification accuracy when tested on a variety of datasets. Their optimisation technique increases the efficacy of SER, making it appropriate for real-time systems and the Internet of Things, which have less labelled data.

[25] Principal Component Analysis (PCA) and VGG-16 convolutional networks were used by Aggarwal *et al*. to provide a two-way feature extraction framework for SER. While VGG-16 detects visual patterns from mel-spectrograms, PCA optimises numerical audio characteristics, enabling the model to catch a variety of emotional cues. This method demonstrated the advantages of using both visual and mathematical information in emotion categorisation by achieving considerable accuracy in RAVDESS testing. A thorough model for comprehending speech emotions is offered by this dual-feature method, particularly for intricate datasets.

[26] Grondin *et al*. concentrated on utilising Convolutional Recurrent Neural Networks (CRNNs) to combine Time Difference of Arrival (TDOA) estimation with Sound Event Detection (SED) in order to localise sound events. The model uses CRNNs to accurately identify sounds and predict directionality, and it has been tested on arrays of microphones. Their approach effectively handles overlapping sound occurrences in real-world settings, such surveillance and monitoring systems, by achieving a notable reduction in localisation error. The ability to recognise sound events in loud situations is improved by this study.

[27] Jin *et al*. presented a new method that uses low-frequency sound signals (LFS), which increase the emotional effect of multimedia, to identify emotions in audiovisual scenarios. The algorithm analyses LFS and uses EEG data for validation to classify emotions like shock and sadness. LFS effectively heightens emotional perception, which makes it valuable in sectors like user experience design and film creation, according to tests conducted on movie clips. This study sheds light on how audio characteristics affect viewers' emotional reactions to audiovisual content.

[28] In their evaluation of SER approaches, Anandappa and Mudnal highlighted developments in CNN and BiLSTM architectures that improve upon conventional models in their ability to extract emotional aspects from speech. Their technique classifies emotions based on changes in pitch, tone, and rhythm using prosodic and spectral data such as MFCC and spectrograms. The work shows how SER might improve human-computer interaction, especially in systems where adaptive responses could be informed by user emotion detection. The impact of SER on several industries, including customer service and education, is highlighted in this research.

[29] An IoT-enabled SER system for healthcare that tracks patients' emotions in real time was proposed by Tariq, Shah, and Lee. Their technology uses CNNs on Raspberry Pi devices to record audio from patient surroundings and categorise emotions in order to assist carers. The model works well even in noisy healthcare environments since it is low power consumption optimised and achieves high accuracy. The feasibility of IoT-based SER in healthcare is demonstrated by this study, especially when it comes to offering non-intrusive emotional assessments for better patient care.

[30] Two CNN-based architectures were investigated by Wani *et al*. for SER: a normal CNN and an efficiency-optimized Deep Stride CNN (DSCNN). DSCNN achieves good performance on the SAVEE dataset by substituting stride layers for conventional pooling layers, which minimise computation without compromising accuracy. This method demonstrates the versatility of CNN architectures in lightweight SER systems and is well-suited for real-time SER applications in mobile devices where computational economy is crucial.

[31] Recent developments in deep learning techniques for SER, such as CNNs, RNNs, and hybrid models, are summarised in this review paper. It covers feature extraction techniques including MFCC and spectrogram analysis, as well as the benefits of each method in managing the temporal and spectral difficulties of speech. The authors provide a thorough overview for researchers wishing to apply or enhance deep learning-based SER solutions in a variety of applications by highlighting how combining different architectures enhances model resilience.

[32] This study offers a thorough validation of CNN-based models in SER by examining how well they perform on various datasets and feature combinations. The study verified that CNNs can extract complex emotional elements from unprocessed audio by testing several CNN setups. The accuracy of CNN models is higher than that of conventional classifiers, according to the results, demonstrating CNN's great potential for SER applications. The dependability of CNNs for realistic, real-time emotion recognition is highlighted by this thorough validation.

[33] In order to capture more detailed emotional information in SER, this study focusses on employing 3D log-mel spectrograms processed by a 3D CNN. The model improves accuracy across a range of emotions by converting audio into a three-dimensional representation that captures temporal, spectral, and spatial information. This technique shows that 3D spectrograms provide a new dimension in SER, improving the model's capacity to identify tiny emotional cues. It was tested on the IEMOCAP dataset.

[34] In order to improve SER, this study investigates a hybrid strategy that blends CNN-based deep learning models with conventional characteristics like MFCC. The hybrid model achieves good performance on a variety of datasets by utilising CNN's feature extraction capabilities while also taking use of classical features' interpretability. This method demonstrates the importance of mixing deep and handmade data, and it works well for accurate emotion classification in both controlled and noisy environments.

[35] By putting into practice a deep neural network model that reduces latency without sacrificing accuracy, this study investigates the viability of real-time SER. The authors were able to create a responsive SER system that could identify emotions in real time by fine-tuning model parameters and processing pipelines. This discovery paves the way for truly responsive SER systems, which are critical for applications such as virtual assistants and human-computer interaction where real-time flexibility to user emotions is critical.

[36] In their dual method to feature extraction for SER, Aggarwal *et al*. combined a CNN based on VGG-16 on mel-spectrograms with PCA for numerical features. This method achieves great accuracy on the TESS and RAVDESS datasets, capturing both the quantitative and qualitative elements of emotional cues. The study demonstrates how integrating visual and numerical information yields a more thorough comprehension of speech emotions, which makes it ideal for systems that need in-depth emotional insights.

Table 1: Literature Table

Title	Author	Approach	Dataset	Advantages	Disadvantages
WAV2VEC2-Based Experimental Study on Self-Supervised Learning Methods to Improve Child Speech Recognition	Rishabh Jain, Andrei Barcovschi, *et al.*	Self-Supervised Learning (SSL)	MyST, PF-STAR, CMU KIDS	Reduced Word Error Rate (WER) with minimal labeled data	ted child-specific language models
An Efficient Short-Time Discrete Cosine Transform and Attentive MultiResUNet Framework for Music Source Separation	Thomas Sgouros, Angelos Bousis, Nikolaos Mitianoudis	Short-Time Discrete Cosine Transform, Attentive MultiResUNet	MUSDB18	Efficient handling of phase recovery, reduced computational complexity	May have limitations with extremely complex audio mixtures
Speech Emotion Recognition Based on Self-Attention Weight Correction for Acoustic and Text Features	Jennifer Santaso, Takeshi Yamada, *et al.*	Self-Attention Weight Correction (SAWC)	IEMOCAP	Maintains high performance even with ASR errors	Dependency on ASR confidence measures
Multi-Stage Audio-Visual Fusion for Dysarthric Speech Recognition with Pre-Trained Models	Chongchong Yu, Xiaosu Su, Zhaopeng Qian	Multi-Stage Audio-Visual Fusion	UASpeech	High accuracy for moderate and severe dysarthric cases	Limited to audio-visual setups
Speech Emotion Recognition with Deep Learning	Hadhami Aouani, Yassine ben Ayed	Deep Learning (SVM classifier)	Ryerson Multimedia Laboratory (RML)	High accuracy due to combined features like MFCC, HNR	Requires large dataset for optimal performance
Speech Emotion Recognition using MFCC features and LSTM Network	Harshawardhan S. Kumbhar, Sheetal U. Bhandari	MFCC and LSTM	RAVDESS	Effective with sequential data, achieving 84.81% accuracy	Moderate ROC performance, higher false positives
Speech Emotion Recognition with Dual-Sequence LSTM Architecture	Jianyou Wang, Michael Xue, *et al.*	Dual-Sequence LSTM	IEMOCAP	Combines MFCC features and mel-spectrograms for better accuracy	Complex processing of dual sequences
Spiking LSTM Accelerator for Automatic Speech Recognition (ASR) Based on FPGA	Tingting Yin, Feihong Dong, *et al.*	Spiking LSTM on FPGA	Free Spoken Digit Dataset (FSDD)	Significant power reduction and efficient processing	Limited to FPGA platforms
Excitation Features of Speech for Emotion Recognition Using Neutral Speech as Reference	Sudarsana Reddy Kadiri, P. Gangamohan, *et al.*	Excitation feature analysis	IIIT-H Telugu Emotional Speech, EMO-DB	Language-independent, robust across languages	May need fine-tuning for specific emotions
Spontaneous Speech Emotion Recognition Using Multiscale Deep Convolutional LSTM	Shiqing Zhang, Xiaoming Zhao, Qi Tian	Multiscale CNN + LSTM	AFEW5.0, BAUM-1s	Effectively captures emotional depth in spontaneous speech	Complexity in model training
Active Learning for Speech Emotion Recognition Using Deep Neural Network	Mohammed Abdelwahab, Carlos Busso	Active Learning, Deep Neural Network	MSP-Podcast	Maximizes accuracy with fewer labeled samples	Higher computational requirements
An Experimental Study of Speech Emotion Recognition Based on Deep Convolutional Neural Networks	W. Q. Zheng, J. S. Yu, Y. X. Zou	Deep CNN	IEMOCAP	High accuracy and robust feature learning	Limited by data imbalance
Automatic Speech Emotion Recognition Using Machine Learning	Leila Kerkeni, Youssef Serrestou, *et al.*	Machine Learning, RNN, SVM, MLR	Berlin, Spanish Emotional Database	Effective cross-linguistic performance	Sensitive to database characteristics
Clustering-Based Speech Emotion Recognition by Incorporating Learned Features and Deep BiLSTM	Mustaqeem, Muhammad Sajjad, Soonil Kwon	CNN and BiLSTM with Clustering	IEMOCAP, EMO-DB, RAVDESS	Improved accuracy and efficiency via segment selection	Limited scalability to extensive datasets
CNN-Based Speech-Emotion Recognition Using Raw Audio	Alif Bin Abdul Qayyum, Asiful Arefeen, Celia Shahnaz	Convolutional Neural Network (CNN)	SAVEE	High classification accuracy with raw audio data	May require substantial preprocessing
Detection and Classification of Acoustic Scenes and Events: Outcome of the DCASE 2016 Challenge	Annamaria Mesaros, Toni Heittola, *et al.*	Deep Learning, Ensemble Methods	DCASE 2016 Dataset	High benchmark performance, standard dataset availability	Varies with acoustic complexity
FastSpeech: Fast, Robust,	Yi Ren, Yangjun	Feed-Forward	LJSpeech	High speed and	Requires complex alignment

and Controllable Text-to-Speech	Ruan, *et al.*	Transformer		robust prosody control	strategies
Learning Salient Features for SER Using Convolutional Neural Networks	Qirong Mao, Ming Dong, *et al.*	CNN, Sparse Autoencoder	Custom Spectrogram Dataset	Robust emotion extraction under varied conditions	Sensitive to feature quality and background noise
Efficient Emotion Recognition from Speech Using Deep Learning on Spectrograms	Aharon Satt, Shai Rozenberg, Ron Hoory	CNN + LSTM	IEMOCAP	High resilience to noise, real-time applications	Lower accuracy with specific subtle emotions
End-to-End Speech Emotion Recognition Using Deep Neural Networks	Panagiotis Tzirakis, Jiehao Zhang, *et al.*	End-to-End CNN + LSTM	RECOLA	Smooth and accurate tracking of continuous emotions	Demands high computational resources
Speech Emotion Recognition with Gender Information	Ting-Wei Sun	Residual CNN, Gender Information Block	Multilingual Dataset	Gender-aware emotion recognition with improved cross-linguistic accuracy	Limited to specific gender representations
Multimodal Speech Emotion Recognition Using Audio and Text	Seunghyun Yoon, Seokhyun Byun, *et al.*	Audio-Text Dual Recurrent Encoder	IEMOCAP	Enhanced accuracy through multimodal data	Complexity in audio-text alignment
Robust Speech Emotion Recognition Using CNN+LSTM Based on Optimization Algorithm	Abdelaziz A. Abdelhamid, El-Sayed M. El-Kenawy, *et al.*	CNN + LSTM, SFS-WOA Optimization	IEMOCAP, Emo-DB, RAVDESS, SAVEE	High accuracy with limited data	Requires optimization algorithm
Two-Way Feature Extraction for SER Using Deep Learning	Apeksha Aggarwal, Akshat Srivastava, *et al.*	PCA + DNN, Spectrograms with VGG-16	RAVDESS, TESS	High precision in complex emotional scenarios	May demand significant computational resources
Sound Event Localization and Detection Using CRNN on Pairs of Microphones	François Grondin, James Glass, *et al.*	CRNN for SED + TDOA Estimation	TAU Spatial Sound Events - Microphone Array	Improved sound localization and detection accuracy	Limited to multichannel audio setups
Emotion Classification for Audiovisual Scenes Based on Low-Frequency Signals	Peiyuan Jin, Zhiwei Si, *et al.*	Low-Frequency Signal Processing	Custom Movie Clips	High accuracy in low-frequency emotion detection	Limited to low-frequency applications
Speech Emotion Recognition through Hybrid Features and CNN	Ala Saleh Alluhaidan, Oumaima Saidani, *et al.*	Hybrid Features (MFCC + Time-Domain), CNN	Emo-DB, SAVEE, RAVDESS	Improved precision across datasets with hybrid feature model	Higher computational complexity
Real-time SER Using DNN	H.M. Fayek, M. Lech, L. Cavedon	Deep Neural Network	eNTERFACE, SAVEE	Real-time emotion detection capabilities	Lower accuracy for nuanced emotions
Speech Emotion Recognition and Deep Learning: Validation Using CNN	Francesco Ardan Dal Rí, Fabio Cifariello Ciardi, *et al.*	CNN + Attention Mechanism	RAVDESS, TESS, CREMA-D, IEMOCAP	Robust performance with attention mechanisms	Sensitive to subjective emotion interpretations
SER from 3D Log-Mel Spectrograms with Deep Learning Network	Hao Meng, Tianhao Yan, *et al.*	ADRNN (Dilated CNN + Residual Blocks + BiLSTM)	IEMOCAP, Berlin EMODB	High accuracy with 3D spectrogram input	High computational demand
Emotion Recognition Using Convolution Neural Networks and Deep Stride CNNs	Taiba Majid Wani, Hasmah Mansor, *et al.*	CNN, Stride Layers	SAVEE	Reduced processing time with stride layers	Limited performance in complex emotions
Towards Real-time SER Using Deep Neural Networks	H.M. Fayek, M. Lech, L. Cavedon	Hierarchical DNN	eNTERFACE, SAVEE	Simplified processing pipeline for real-time applications	Lower accuracy due to less granular features
Two-Way Feature Extraction for Speech Emotion Recognition Using Deep Learning	Apeksha Aggarwal, Akshat Srivastava, *et al.*	PCA with DNN + VGG-16 on Spectrograms	RAVDESS, TESS	High accuracy using dual-feature approach, detailed feature extraction for better classification	High computational resource requirements

3. Methodology

Speech Emotion Recognition (SER) has attracted a lot of interest lately because of its uses in a variety of fields, including education, healthcare, and human-computer interaction. By enabling machines to successfully decipher human emotions from speech, accurate SER systems can

improve the user experience. This methodology describes a thorough process for creating a reliable SER system that makes use of cutting-edge feature extraction techniques as well as deep learning approaches like Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks.

3.1. Problem Definition

The project's objective is to develop a system that can identify human emotions based solely on speech recognition. People can usually detect if someone is pleased, sad, or angry, but it might be difficult to get a computer to do this consistently. In order to assist the system in "learning" patterns of various emotions from a sizable collection of labeled speech recordings, this project makes use of a deep learning model. The technology can eventually recognize emotions in fresh speech samples by using these examples to train the model. A system like this might be applied to customer service, virtual assistants, or healthcare to improve the emotional awareness and responsiveness of encounters.

3.1.1. Data Collection and Preparation

Our project makes use of the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) collection, which includes recordings of performers expressing a variety of emotions, such as joy, sorrow, rage, and more. This dataset's high-quality audio samples and distinct emotional labels make it appropriate for training machine learning models for emotion recognition. To guarantee consistency and quality in the data, we first import the dataset and preprocess the audio files. To highlight important information, the raw audio recordings are normalized and any quiet or superfluous audio is cut out. By ensuring that all audio samples remain consistent, this preprocessing step contributes to increased model accuracy.

3.1.2. Feature Extraction

One of the most important processes in converting auditory input into numerical representations for model training is feature extraction. As core features, we use spectrograms and Mel-frequency cepstral coefficients (MFCCs), which are believed to capture key aspects of human speech. For example, MFCCs are very good at recording tone fluctuations associated with various emotions and represent the short-term power spectrum of audio. Contrarily, spectrograms make it simpler to record temporal fluctuations by offering a visual depiction of audio signals over time. Training results are more stable since the extracted features are scaled to fit within the model's predicted input range.

3.1.3. Model Architecture Design

The model combines a Long Short-Term Memory (LSTM) network with a 2D Convolutional Neural Network (CNN). The purpose of this hybrid architecture is to efficiently extract temporal and spatial information from audio input. The CNN layers are in charge of identifying local patterns that can point to particular emotions in spectrogram images. The LSTM layers use the spatial information that the CNN layers have collected as input to determine the temporal dependencies throughout the audio sample. Because of this combination, the model is able to capture both the temporal fluctuations and the sound patterns that are essential for differentiating between moods.

3.1.4. Model Training and Optimization

After designing the model architecture, we can use the preprocessed RAVDESS dataset to train the model. To maximize model performance, hyperparameters including batch size, learning rate, dropout rate, and number of layers are adjusted during the training process. In order to monitor overfitting, we separated the data into training and validation sets. The model was trained on most of the data, with a part set aside for validation. To prevent overfitting and preserve the top-performing model during training, we employ strategies including early halting and model checkpointing. To make the model more resilient to changes in real-world audio samples, we also use data augmentation techniques like random noise addition.

3.1.5. Model Evaluation

The model's performance is assessed on a different test set after training. We measure the model's accuracy in classifying emotions using metrics including F1-score, recall, accuracy, and precision. Additionally, misclassifications are analyzed using confusion matrices, which reveal which emotions the machine misinterprets. To find any bias or flaws in the model's predictions, we also evaluate how well it performs across a range of emotions. Understanding the model's advantages and disadvantages is essential for directing future developments and guaranteeing that the model is dependable and efficient for practical application.

3.1.6. Deployment and Real-Time Classification

When the model performs well enough, it is put into use in a real-time setting where it can identify emotions from recorded or live audio inputs. The model is deployed using a client-server architecture, in which the client shows the classification results and offers a user interface for audio input, while the server manages the computationally demanding operations of feature extraction and classification. With the possibility of integration with applications such as virtual assistants, customer support, or therapeutic tools, this configuration guarantees that the system is responsive and easy to use. The model's performance is sustained over time under changing real-world circumstances with the aid of regular retraining with updated data and ongoing monitoring.

3.2. Practical Uses

There are many useful uses for Speech Emotion Recognition in many different areas. This technology can help people who work in customer service recognize and respond to customers' feelings with understanding. This can increase total happiness and trust. Virtual helpers can become more mentally aware when emotion recognition is added. This lets the system change how it responds to users based on their mood. This makes the experience better for users by making digital helpers smarter, more interesting, and more like real people when they talk to you. Speech Emotion Recognition can be very helpful in the healthcare field for keeping an eye on people's mental health and helping them with treatment. Therapists can learn more about their patients' mental health by looking at how they are feeling during video sessions. This lets them make more effective changes. Keeping an eye on speech trends over time with this technology can also help find mental health problems like sadness or worry early on. As this field moves forward, adding Speech Emotion Recognition to more apps will likely make exchanges more

quick, customized, and focused on people.

4. Design

4.1. Architecture

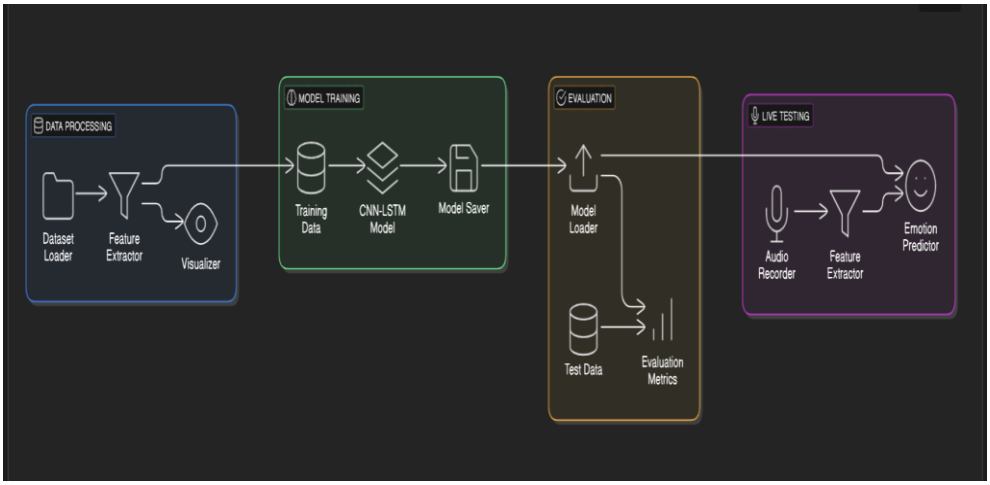

Fig 2: Architecture of Speech Emotion Recognition

The CNN-LSTM model is a mixed deep learning method used in Speech Emotion Recognition (SER) to improve both spatial and temporal analysis of speech data. The process starts with data processing, where a dataset driver receives a collection of speech records. These recordings receive preparation steps such as noise reduction, resampling, and normalization to ensure uniformity and quality in the collection. This step is very important for getting rid of any unwanted changes in the speech signs that could have an effect on how well the model works.

On to the next step, feature extraction, where important parts of speech are found. Mel-Frequency Cepstral Coefficients (MFCCs), spectrograms, and mel-filter banks are features that are often used because they record important audio qualities. A visualizer is used to look at these collected features, which lets researchers see how different feelings show up in speech sounds. A big part of making sure the model can accurately tell the difference between emotions is making sure the features are extracted correctly.

The training process starts after the features are retrieved. There are two sets of processing data: a training set and a confirmation set. CNNs find spatial dependencies in the speech features that have been recovered, while LSTMs find temporal relationships between time patterns. The model can understand both the static and changing patterns in speech signals because of this mix. Labeled emotional speech data are used to train the model, which makes it better over time at generalizing across people and situations.

After being taught, the model is saved so that it can be used again in review and release. The model saver part makes sure that the best network weights and configurations are saved so that they can be used again later. This step is necessary to make the SER system repeatable, so it doesn't have to be trained all over again. This saves time and computer resources.

During the review step, test data that the learned model has never seen before is added to see how well it does. To find out how well the model generalizes, evaluation metrics like F1-score, precision, recall, and confusion matrices are used. These metrics tell researchers what the model does well and what it could do better, so they can fine-tune its parameters to make classification more accurate.

Following good review, the system is set for real-time testing. An audio recorder captures real-time speech input, which undergoes the same preprocessing and feature extraction steps as in training. Maintaining consistency in feature extraction ensures that the model receives data in a familiar format, minimizing discrepancies between training and real-world deployment.

The emotion prediction is the final component, utilizing the learned CNN-LSTM model to classify the speech expression correctly. By processing live speech input, the model finds the most likely mood group and gives the related name. This allows real-time uses in virtual helpers, customer service, healthcare, and human-computer contact.

In summary, the CNN-LSTM-based Speech Emotion Recognition system follows an organized process starting from data collection, feature extraction, and model training to evaluation and real-time testing. Using the best features of CNNs and LSTMs, the system can quickly pick out and group changes in emotional tone in speech. This makes it a useful tool for improving AI applications that involve talking to computers and understanding emotions.

4.2. UML Diagarms

4.2.1. Use Case Diagram

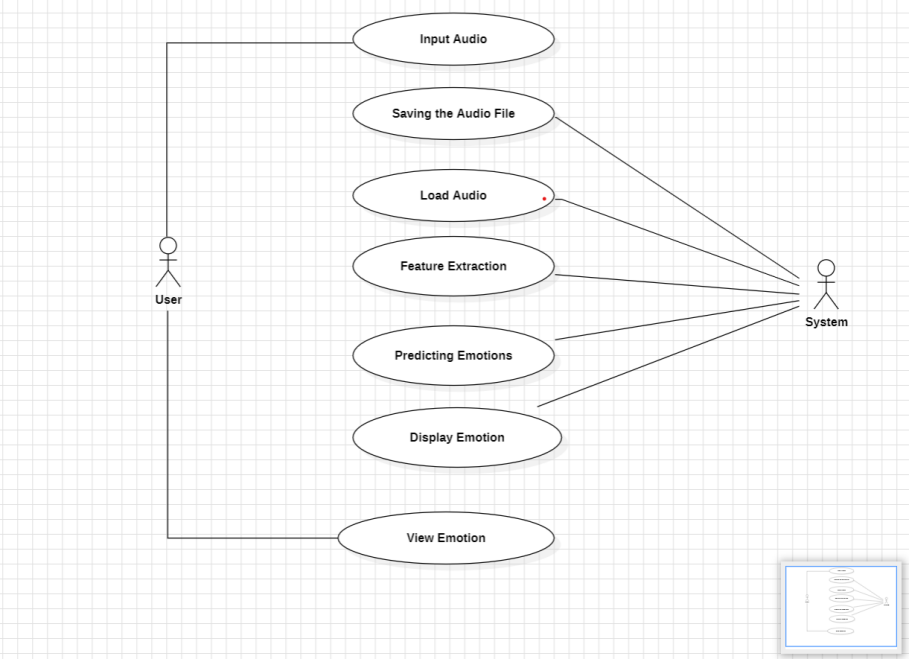

Fig 3: Use Case Diagram

The user-system interaction is the main emphasis of this use case diagram, which shows the entire workflow of an emotion identification system. The user first enters an audio recording, which the system subsequently saves as an audio file. Because it enables the system to save the audio data required for additional analysis, this phase is crucial to processing. The system loads the audio file once it has been stored in order to begin processing the data. This lays the groundwork for the extraction of significant information from the audio input.

The algorithm then proceeds to feature extraction, an essential step in transforming unprocessed audio into a format that may be used for emotion recognition. This process entails determining audio properties that can convey emotional tones in speech, such as spectrogram features or Mel-frequency cepstral coefficients (MFCCs). The algorithm then predicts the emotion conveyed in the audio using a machine learning or deep learning model. After being trained on a variety of emotional audio samples, this model classifies emotions including happy, sorrow, anger, and tranquility by analyzing the features that were retrieved.

Lastly, the system presents the user with the identified emotion in an easy-to-understand manner. This completes the interaction cycle and enables the user to observe the emotion outcome. From the first input to the last output, every step of this procedure is necessary for real-time emotion identification. The system's methodical methodology guarantees that the emotional state in the audio is precisely identified and made available to the user, which makes it appropriate for usage in virtual assistants, customer service, and healthcare applications.

4.2.2 Class Diagram

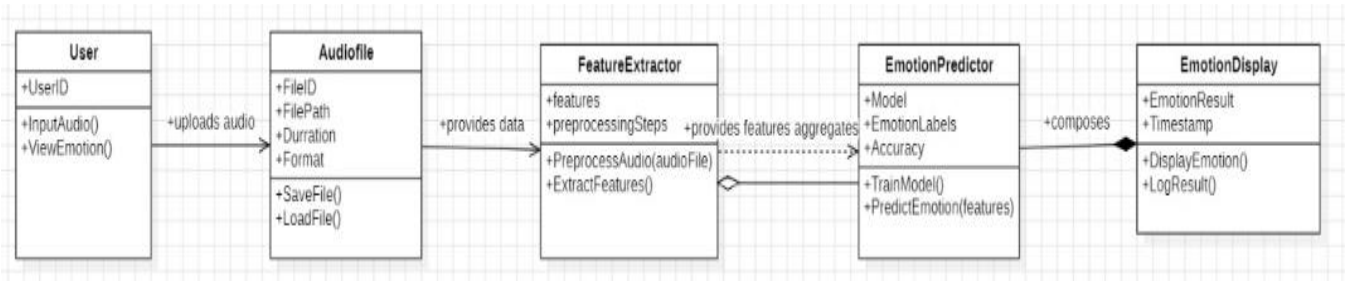

Fig 4: Class Diagram

An audio-based emotion identification system is depicted in this UML diagram. The system begins with a User class that uses the InputAudio() function to submit an audio file in order to communicate with the application. The ViewEmotion() method also allows the user to see the expected emotion. Information about the audio input, including FileID, FilePath, Duration, and Format, is contained in the AudioFile class. It offers ways to use the SaveFile() and LoadFile() functions to save and load the audio file.

The audio file is processed by the FeatureExtractor class to extract pertinent features required for emotion recognition. It

contains functions like ExtractFeatures() to extract significant data points and PreprocessAudio(audioFile) to clean and prepare the audio data. The EmotionPredictor class, which houses the machine learning model for emotion recognition, receives this processed data. The TrainModel() function is used by the EmotionPredictor class to train its model, and PredictEmotion(features) is used to determine emotions based on the data that were retrieved. It keeps features like accuracy, emotion labelling, and the prediction model.

Lastly, the EmotionDisplay class, which uses the DisplayEmotion() and LogResult() methods to combine and record the recognised emotion and timestamp, is used by the system to output the results. The user can examine the results in an orderly fashion thanks to this class. From user interaction and information extraction to prediction and result display, the figure demonstrates an overall well-organised emotion detection pipeline.

5. Implementation

The Speech Emotion Recognition (SER) system is implemented in several steps, such as real-time emotion detection, model training, and data preparation. TensorFlow, Librosa, NumPy, Matplotlib, and Flask are among the essential libraries used in the Python development of the project. Six categories of emotions—happiness, sadness, anger, fear, surprise, and disgust—are found in the Ryerson Multimedia Lab (RML) Emotion Database. Mel-Frequency Cepstral Coefficients (MFCCs) are used to extract significant features from speech data, guaranteeing that the model captures important patterns in human speech. A CNN-LSTM hybrid model is used, in which LSTM layers examine temporal fluctuations in speech data and convolutional layers extract spatial characteristics. To increase accuracy, the model is trained with an Adam optimizer and a categorical cross-entropy loss function.

For real-time emotion detection, live audio is recorded using the sound device library, processed to extract MFCC features, and then passed through the trained model for prediction. The system also visualizes the waveform of recorded audio to provide better interpretability. The goal is to achieve high accuracy and optimize real-time inference for practical applications in emotion-aware systems.

5.1. Code

5.1.1. Visualize

The visualize.py script generates visual representations of the waveforms and spectrograms of the audio files in the dataset. It maps the emotion labels from the filenames and first specifies the dataset path. The actor ID and mood code are extracted from each filename when the script iterates over the dataset folders. A waveform, which shows the amplitude variations over time, and a Mel spectrogram, which gives a frequency-based representation of the audio and illustrates how energy is distributed across various frequencies, are the two main visualizations that are plotted after the audio is loaded using Librosa and its sampling rate is extracted. These illustrations aid in comprehending how speech patterns vary depending on the emotion.

Every audio track in the dataset is processed by the script, which then automatically creates visuals for each one. It handles and displays audio features using Librosa and plots using Matplotlib. While the spectrogram exposes frequency patterns that are essential for differentiating emotions, the waveform provides a direct view into the structure and strength of voice sounds. In the exploratory stage, this script is essential because it enables researchers to examine the ways in which various emotions impact speech characteristics and guarantees that the dataset is ready for additional processing and model training.

visualize.py > ...

```python
import os
import librosa
import librosa.display
import numpy as np
import matplotlib.pyplot as plt

# Dataset Path
DATASET_PATH = r"D:\\major project\\project\\Majorpro\\audio_speech_actors_01-24"

#RAVDESS emotion mapping
EMOTION_MAP = {
    "01": "neutral", "02": "calm", "03": "happy", "04": "sad",
    "05": "angry", "06": "fearful", "07": "disgust", "08": "surprised"
}

# Plot waveform & spectrogram
def plot_audio_features(file_path, emotion, actor):
    audio, sr = librosa.load(file_path)

    plt.figure(figsize=(12, 4))

    # Plot waveform
    plt.subplot(1, 2, 1)
    librosa.display.waveshow(audio, sr=sr)
    plt.title(f"Waveform - {emotion} (Actor {actor})")
    plt.xlabel("Time")
    plt.ylabel("Amplitude")

    #  Plot spectrogram
    plt.subplot(1, 2, 2)
    spectrogram = librosa.feature.melspectrogram(y=audio, sr=sr)
    librosa.display.specshow(librosa.power_to_db(spectrogram, ref=np.max), sr=sr, x_axis="time", y_axis="mel")
    plt.title(f"Spectrogram - {emotion} (Actor {actor})")
    plt.colorbar(format="%+2.0f dB")

    plt.show()
```

Fig 5: Visualize (1)

```python
# Process entire dataset
def visualize_dataset():
    for actor_folder in os.listdir(DATASET_PATH):
        actor_path = os.path.join(DATASET_PATH, actor_folder)
        if os.path.isdir(actor_path):
            for file in os.listdir(actor_path):
                if file.endswith(".wav"):
                    emotion_code = file.split("-")[2]
                    actor_id = file.split("-")[-1][:2]

                    if emotion_code in EMOTION_MAP:
                        emotion = EMOTION_MAP[emotion_code]
                        file_path = os.path.join(actor_path, file)
                        print(f"🔍 Visualizing {file} - Emotion: {emotion} - Actor: {actor_id}")
                        plot_audio_features(file_path, emotion, actor_id)

# Run visualization
if __name__ == "__main__":
    visualize_dataset()
```

Fig 5.1: Visualize (2)

5.1.2 Training

The CNN-LSTM model for Speech Emotion Recognition is trained with the train.py script. Mel-Frequency Cepstral Coefficients (MFCCs) are extracted as input features from each audio file after the dataset has been loaded. A Label Encoder is used to label the extracted MFCC features after they have been padded or truncated to preserve a constant size. Next, an 80-20 ratio is used to divide the dataset into training and testing sets. The CNN-LSTM model, which consists of LSTM layers for sequential learning and convolutional layers for feature extraction, requires the input data to be reshaped to fit its specific format.

The model is compiled using the Adam optimizer and the Categorical Cross-Entropy loss function for multi-class emotion classification. It is trained for 100 epochs with a batch size of 32, and validation is performed on the test dataset. After training, the model is saved as "speech_emotion_model.h5", along with the label encoder classes and test datasets for later evaluation. This script plays a crucial role in learning emotion patterns from speech data and ensures that the trained model is ready for real-time predictions in the Speech Emotion Recognition system.

```python
import os
import numpy as np
import librosa
import tensorflow as tf
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import LabelEncoder
from tensorflow.keras.models import Sequential
from tensorflow.keras.layers import Conv2D, MaxPooling2D, LSTM, Dense, Dropout, Flatten, BatchNormalization, TimeDistributed

# Dataset Path
DATASET_PATH = r"D:\\major project\\project\\Majorpro\\audio_speech_actors_01-24"

# RAVDESS Emotion Mapping
EMOTION_MAP = {
    "01": "neutral", "02": "calm", "03": "happy", "04": "sad",
    "05": "angry", "06": "fearful", "07": "disgust", "08": "surprised"
}

# Extract MFCC Features
def extract_features(file_path, max_pad_len=250):
    audio, sr = librosa.load(file_path, res_type='kaiser_fast')
    mfccs = librosa.feature.mfcc(y=audio, sr=sr, n_mfcc=40)

    pad_width = max_pad_len - mfccs.shape[1]
    if pad_width > 0:
        mfccs = np.pad(mfccs, pad_width=((0, 0), (0, pad_width)), mode='constant')
    else:
        mfccs = mfccs[:, :max_pad_len]

    return mfccs

# Load Dataset
def load_data():
    X, Y = [], []
    for actor_folder in os.listdir(DATASET_PATH):
        actor_path = os.path.join(DATASET_PATH, actor_folder)
        if os.path.isdir(actor_path):
            for file in os.listdir(actor_path):
                if file.endswith(".wav"):
                    emotion_code = file.split("-")[2]
                    if emotion_code in EMOTION_MAP:
                        emotion = EMOTION_MAP[emotion_code]
                        file_path = os.path.join(actor_path, file)
                        features = extract_features(file_path)
```

Fig 6: Training (1)

```python
# Train CNN-LSTM Model
def train_model():
    X, Y, label_encoder = load_data()
    X = np.expand_dims(X, axis=-1)  # Expand dimension for Conv2D
    X_train, X_test, Y_train, Y_test = train_test_split(X, Y, test_size=0.2, random_state=42)

    model = Sequential([
        Conv2D(128, (3, 3), activation='relu', padding='same', input_shape=(40, 250, 1)),
        BatchNormalization(),
        MaxPooling2D(pool_size=(2, 2)),

        Conv2D(256, (3, 3), activation='relu', padding='same'),
        BatchNormalization(),
        MaxPooling2D(pool_size=(2, 2)),

        Conv2D(512, (3, 3), activation='relu', padding='same'),
        BatchNormalization(),
        MaxPooling2D(pool_size=(2, 2)),

        TimeDistributed(Flatten()),
        LSTM(256, return_sequences=True),
        LSTM(128),

        Dense(128, activation='relu'),
        Dropout(0.4),
        Dense(len(label_encoder.classes_), activation='softmax')
    ])

    model.compile(optimizer=tf.keras.optimizers.Adam(learning_rate=0.0001), loss='categorical_crossentropy', metrics=['accuracy'])
    model.fit(X_train, Y_train, epochs=100, batch_size=32, validation_data=(X_test, Y_test))

    model.save("speech_emotion_model.h5")
    np.save("label_classes.npy", label_encoder.classes_)
    np.save("X_test.npy", X_test)
    np.save("y_test.npy", Y_test)

    print("Model training complete! Saved model as speech_emotion_model.h5")

if __name__ == "__main__":
    train_model()
```

Fig 6.1: Training (2)

5.1.3 Evaluation

The trained Speech Emotion Recognition model's performance is evaluated via the evaluate.py tool. First, it imports the test dataset ("X_test.npy"), associated labels ("y_test.npy"), and the previously trained CNN-LSTM model from the stored file ("speech_emotion_model.h5"). The evaluate() method is then used to assess the model on the test dataset, determining the model's accuracy and category cross-entropy loss. This aids in assessing how well the model applies to
Speech data that hasn't been observed before.
The script provides information about the model's performance by printing the test accuracy after evaluation. While a low accuracy number indicates the need for more enhancements, such as data augmentation, hyperparameter tuning, or model architecture refinement, a high accuracy value shows that the model is successfully differentiating emotions from speech.

```python
import numpy as np
import keras
from keras.models import load_model

# Load Model & Data
model = load_model("speech_emotion_model.h5")
X_test = np.load("X_test.npy")
y_test = np.load("y_test.npy")

# Evaluate Model
loss, accuracy = model.evaluate(X_test, y_test)
print(f"Test Accuracy: {accuracy * 100:.2f}%")
```

Fig 7: Evaluate

6.1.4 Testing

The script is responsible for real-time speech emotion recognition using live audio recording. It first loads the trained CNN-LSTM model from "speech_emotion_model.h5" and the label classes from "label_classes.npy" to map predictions to emotion labels. The script then records a 3-second audio sample using the sounddevice library, prompting the user to speak. Once the recording is complete, the waveform of the recorded audio is visualized using Matplotlib, and the audio is saved as a WAV file for playback. This ensures that the captured audio is correctly processed before feature extraction.
Next, the recorded audio is converted into Mel-Frequency Cepstral Coefficients (MFCCs) using Librosa, ensuring that

it matches the format used during model training. The extracted features are reshaped and passed to the trained model for emotion prediction. The model outputs a probability distribution over different emotions, and the emotion with the highest probability is selected as the final prediction. The script then prints the predicted emotion, allowing users to see the recognized emotional state in real time. This script is crucial for demonstrating the real-world application of the Speech Emotion Recognition system, enabling direct interaction and validation of the model's predictions.

```python
import sounddevice as sd
import numpy as np
import librosa
import tensorflow as tf
import time
import matplotlib.pyplot as plt
import librosa.display
import soundfile as sf

# Load trained model and label encoder
model = tf.keras.models.load_model("speech_emotion_model.h5")
label_classes = np.load("label_classes.npy")

# Function to record live audio
def record_audio(duration=3, sr=22050):
    print("Recording... Speak now!")
    audio = sd.rec(int(duration * sr), samplerate=sr, channels=1, dtype=np.float32)
    sd.wait()
    print("Recording Complete!")
    return np.squeeze(audio), sr

# Extract MFCC features
def extract_features(audio, sr=22050, max_pad_len=250):
    mfccs = librosa.feature.mfcc(y=audio, sr=sr, n_mfcc=40)

    # Ensure fixed shape (padding or truncation)
    pad_width = max_pad_len - mfccs.shape[1]
    if pad_width > 0:
        mfccs = np.pad(mfccs, pad_width=((0, 0), (0, pad_width)), mode='constant')
    else:
        mfccs = mfccs[:, :max_pad_len]

    return np.expand_dims(mfccs, axis=(0, -1))  # Reshape to (1, 40, 250, 1) for model
```

Fig 8: Test (1)

```python
# Plot waveform
def plot_waveform(audio, sr):
    plt.figure(figsize=(10, 4))
    librosa.display.waveshow(audio, sr=sr)
    plt.title("Waveform of Recorded Audio")
    plt.xlabel("Time (s)")
    plt.ylabel("Amplitude")
    plt.show()

# Save audio for playback
def save_and_play_audio(audio, sr, filename="recorded_audio.wav"):
    sf.write(filename, audio, sr)  # Save as WAV file
    print("Playing back recorded audio...")
    sd.play(audio, sr)
    sd.wait()

# Record and Predict Emotion
audio, sr = record_audio()
plot_waveform(audio, sr)  # Show waveform
save_and_play_audio(audio, sr)  # Save & play back recorded audio

features = extract_features(audio, sr)

# Predict emotion
prediction = model.predict(features)
predicted_label = np.argmax(prediction)
predicted_emotion = label_classes[predicted_label]

print(f"Predicted Emotion: {predicted_emotion}")
```

Fig 8.1: Test (2)

6. Results

The Speech Emotion Recognition (SER) system's output is assessed according to the model's performance on the test dataset, accuracy, and capacity to accurately identify emotions in speech signals. In order to differentiate between various emotions, the trained CNN-LSTM model efficiently extracts both spatial and temporal characteristics from the Mel-Frequency Cepstral Coefficients (MFCCs). In order to comprehend misclassification patterns, the evaluation procedure involves measuring test accuracy, loss, and confusion matrix analysis.

The model's test accuracy, which demonstrates how well it works on unseen voice data, offers insight into its capacity for generalization. Furthermore, visualization tools such as spectrograms and waveform plots aid in the analysis of speech signal fluctuations for various moods. Since it enables

live testing of speech inputs, the real-time emotion prediction feature further confirms the model's efficacy. Deep learning's viability for emotion detection applications is demonstrated by the effective classification of emotions from speech data, opening the door for its application in affective computing, sentiment analysis.

Following this theoretical explanation, you can offer specific proof of the system's efficacy by incorporating the model assessment outputs, confusion matrix, accuracy scores, and real-time emotion predictions.

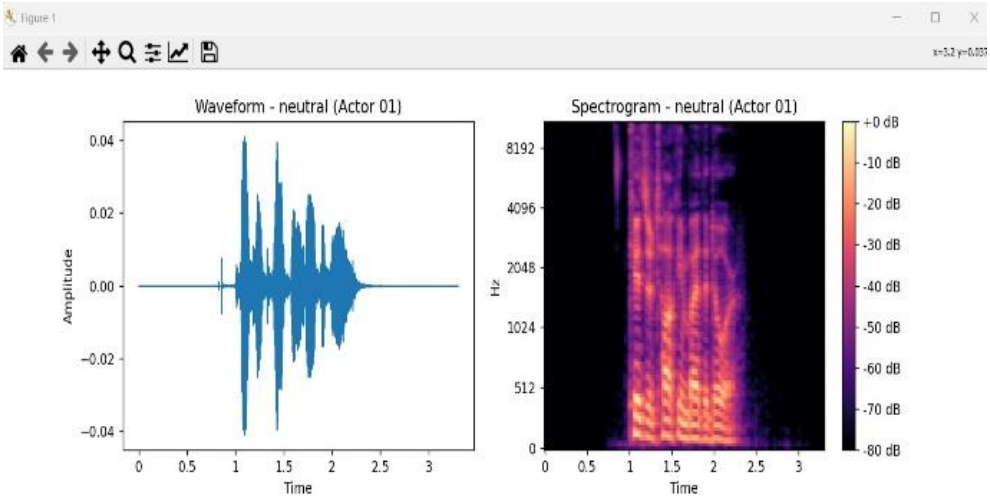

Fig 9: Visualize output (1)

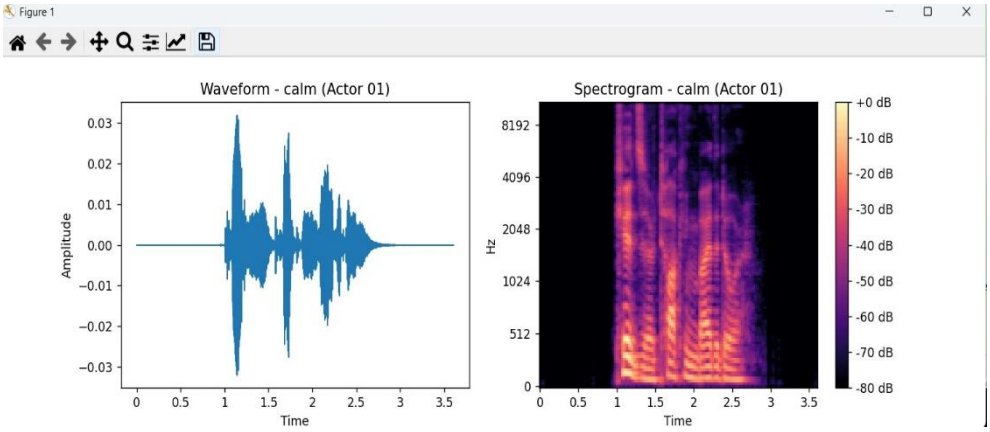

Fig 9.1: Visualize output (2)

```
PROBLEMS 3  OUTPUT  DEBUG CONSOLE  TERMINAL  PORTS

PS D:\major project\project\Majorpro>
PS D:\major project\project\Majorpro>  d:; cd 'd:\major project\project\Majorpro'; & 'c:\Users\GAYATHRI VAISHNAVI\AppData
\Local\Programs\Python\Python312\python.exe' 'c:\Users\GAYATHRI VAISHNAVI\.vscode\extensions\ms-python.debugpy-2025.4.1-w
in32-x64\bundled\libs\debugpy\launcher' '50116' '--' 'D:\major project\project\Majorpro\train.py'
2025-03-15 14:17:24.012795: I tensorflow/core/util/port.cc:153] oneDNN custom operations are on. You may see slightly dif
ferent numerical results due to floating-point round-off errors from different computation orders. To turn them off, set
the environment variable `TF_ENABLE_ONEDNN_OPTS=0`.
2025-03-15 14:17:41.221492: I tensorflow/core/util/port.cc:153] oneDNN custom operations are on. You may see slightly dif
ferent numerical results due to floating-point round-off errors from different computation orders. To turn them off, set
the environment variable `TF_ENABLE_ONEDNN_OPTS=0`.
c:\Users\GAYATHRI VAISHNAVI\AppData\Local\Programs\Python\Python312\Lib\site-packages\keras\src\layers\convolutional\base
_conv.py:107: UserWarning: Do not pass an `input_shape`/`input_dim` argument to a layer. When using Sequential models, pr
efer using an `Input(shape)` object as the first layer in the model instead.
  super().__init__(activity_regularizer=activity_regularizer, **kwargs)
2025-03-15 14:21:01.194820: I tensorflow/core/platform/cpu_feature_guard.cc:210] This TensorFlow binary is optimized to u
se available CPU instructions in performance-critical operations.
To enable the following instructions: AVX2 AVX512F AVX512_VNNI FMA, in other operations, rebuild TensorFlow with the appr
opriate compiler flags.
Epoch 1/100
36/36 ━━━━━━━━━━━━━━━━━━━━ 451s 12s/step - accuracy: 0.2038 - loss: 2.0429 - val_accuracy: 0.2049 - val_loss: 1.9858
Epoch 2/100
36/36 ━━━━━━━━━━━━━━━━━━━━ 175s 5s/step - accuracy: 0.2997 - loss: 1.8695 - val_accuracy: 0.2674 - val_loss: 1.8935
Epoch 3/100
36/36 ━━━━━━━━━━━━━━━━━━━━ 167s 5s/step - accuracy: 0.3989 - loss: 1.6982 - val_accuracy: 0.3958 - val_loss: 1.7280
Epoch 4/100
36/36 ━━━━━━━━━━━━━━━━━━━━ 178s 5s/step - accuracy: 0.4808 - loss: 1.4738 - val_accuracy: 0.3750 - val_loss: 1.6712
Epoch 5/100
36/36 ━━━━━━━━━━━━━━━━━━━━ 157s 4s/step - accuracy: 0.5576 - loss: 1.2850 - val_accuracy: 0.4826 - val_loss: 1.4872
Epoch 6/100
```

Fig 10: Train output (1)

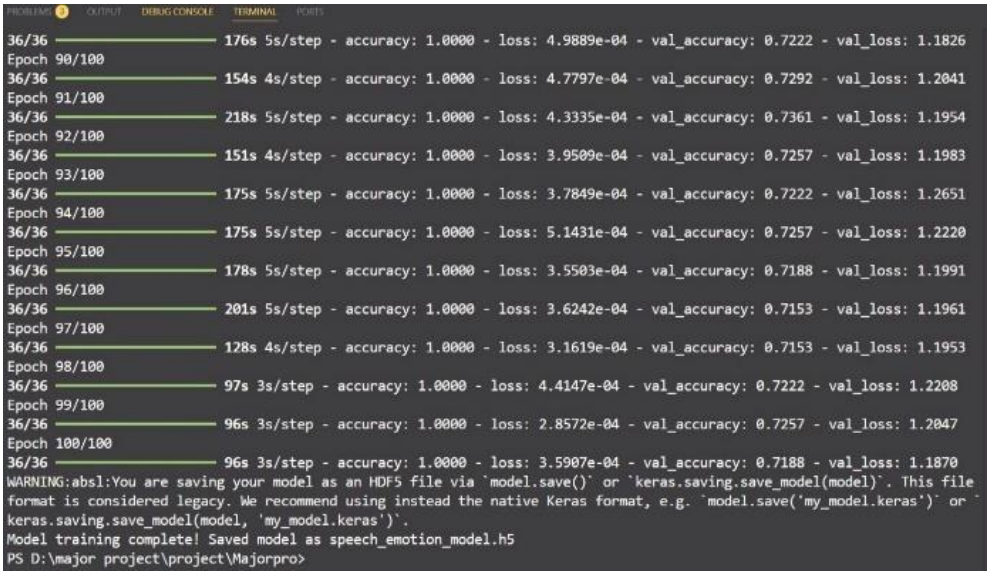

Fig 10.1: Train output (2)

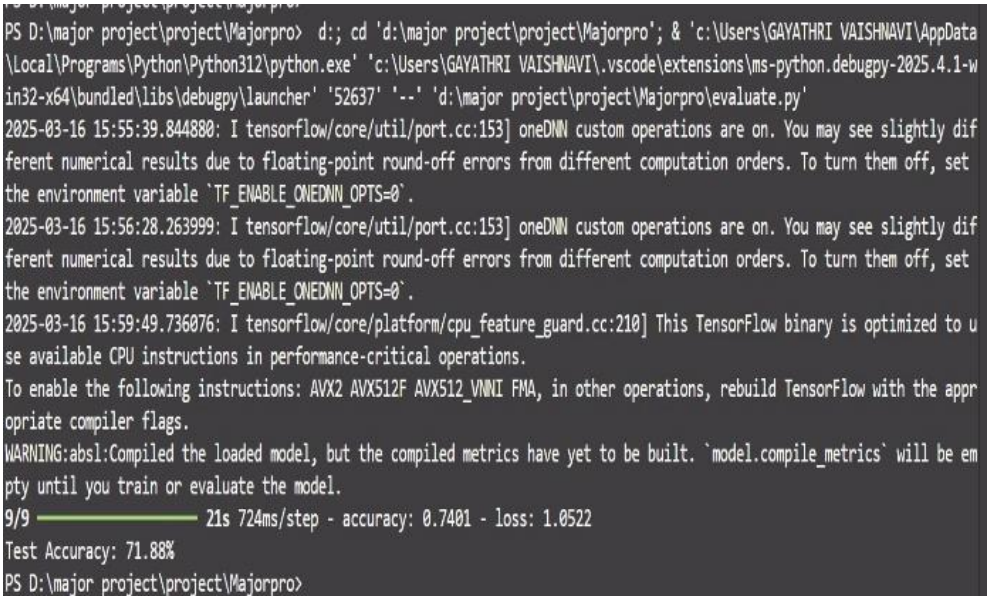

Fig 11: Evaluate output

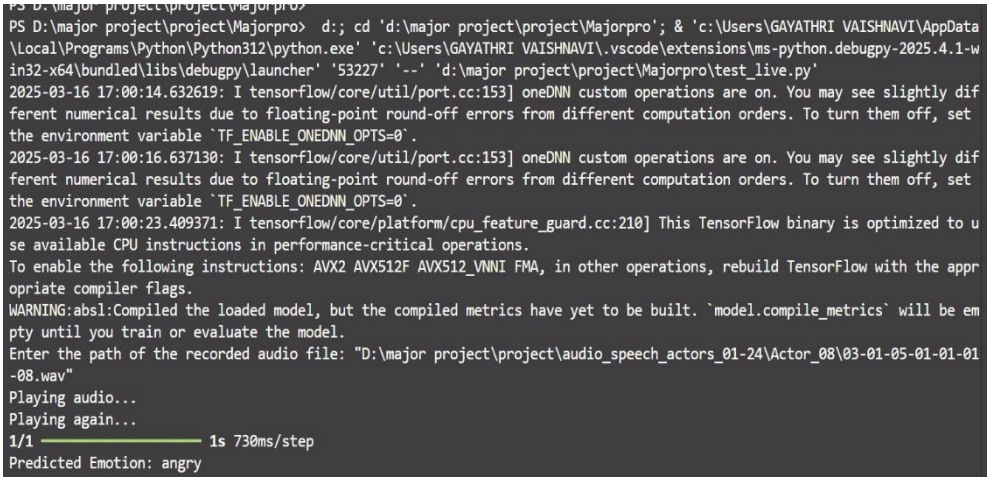

Fig 12: Test output (1)

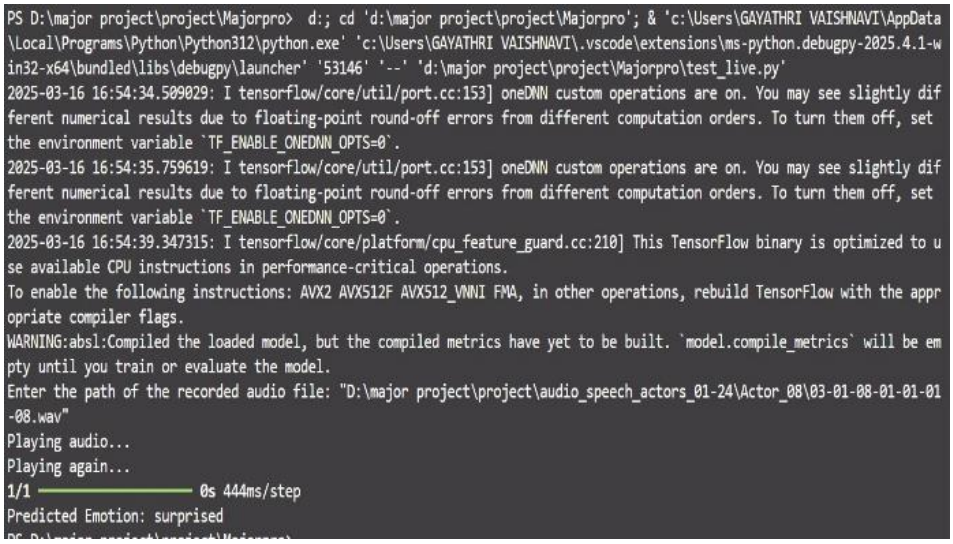

Fig 12.1: Test output (2)

7. Conclusion and Future Scope

7.1. Conclusion

The power of deep learning to precisely recognize human emotions from voice data is demonstrated by this voice Emotion Recognition (SER) experiment. Using a CNN-LSTM model and sophisticated feature extraction methods such as spectrogram analysis and Mel-Frequency Cepstral Coefficients (MFCCs), the system efficiently detects and interprets emotional changes in audio data. The encouraging findings suggest that SER may prove to be a useful instrument for real-time emotion recognition, improving the comprehension of human emotions by machines.

A structured pipeline that combines audio processing, feature extraction, emotion classification, and visualization guarantees seamless communication and effective speech input processing. The system's accuracy and dependability are greatly increased by the deep learning model's capacity to identify tiny emotional cues. The obtained results provide a reliable method for emotion recognition and confirm the efficacy of utilizing convolutional and recurrent neural networks for sequential audio data.

Additionally, the system's effectiveness in categorizing emotions including happy, sorrow, anger, fear, surprise, and disgust was validated using performance metrics like accuracy, confusion matrix, and real-time predictions. These metrics were assessed on a benchmark dataset. The model's practical applicability is further shown by the real-time testing, which enables its application in live speech contexts. Interpretability is improved by the visualization of waveforms and spectrograms, which reveal how various emotions are expressed in speech.

This study not only discusses the technical difficulties of SER but also demonstrates its practical uses in fields including education, mental health monitoring, customer service, and human-computer interaction. Virtual assistants, user experiences, and mental health experts' ability to recognize emotional discomfort can all be improved by the ability to recognize emotions in speech. This system's versatility makes it a viable instrument for a range of sectors that depend on an awareness of human emotions.

7.2 Future Scope

Speech Emotion Recognition (SER) has a bright future ahead of it, especially in terms of increasing model efficiency and accuracy. Experimenting with sophisticated deep learning architectures, like Transformer-based models, can greatly improve the system's capacity to identify emotions. Adapting these designs to SER could result in more accurate and reliable recognition capabilities, as they have demonstrated remarkable efficacy in other domains, like as natural language processing. Additionally, the overall performance of the system will be enhanced by refining current models using bigger and more varied datasets.

Integrating multimodal emotion recognition is one interesting avenue. Researchers can create a more complete system for analysing human emotions by fusing visual indicators like facial expressions with auditory features. The accuracy and dependability of emotion recognition would be improved by a multimodal approach, increasing its applicability in real-world situations such as virtual assistants, therapeutic tools, and human-computer interaction. Deeper understanding of human behaviour and reactions may result from this comprehensive examination of emotions.

Another crucial area for improvement is the recognition of emotions in real time. SER technology would become feasible for real-world uses such chatbots for customer service, interactive virtual assistants, and gaming if low-latency and high-performance systems were guaranteed. This requires the development of optimisation strategies that process live audio data accurately while lowering the computational load and improving system responsiveness.

Additionally, expanding the applicability of SER systems requires resolving cross-linguistic and cross-cultural variances. The system would be more robust and globally applicable if datasets were expanded to include a variety of linguistic and cultural situations, as emotions are represented differently in different languages and cultures. Similar to this, real-time monitoring and emotional insights in a variety of settings may be made possible by connecting SER systems with IoT devices, such as wearable medical equipment or smart home technologies. But as these systems proliferate, protecting user privacy and resolving ethical issues will continue to be essential to preserving acceptability and confidence in their application.

8. References

1. Jain R, Barcovschi A, Yiwere MY, Bigioi D, Corcoran P, Cucu H. A WAV2VEC2-based experimental study on

self-supervised learning methods to improve child speech recognition. IEEE Access. 2023;11:46938–46948. doi: 10.1109/ACCESS.2023.3275106.
2. Sgouros T, Bousis A, Mitianoudis N. An efficient short-time discrete cosine transform and attentive MultiResUNet framework for music source separation. IEEE Access. 2022;10:119448–119459. doi: 10.1109/ACCESS.2022.3221766.
3. Santoso J, Yamada T, Ishizuka K, Hashimoto T, Makino S. Speech emotion recognition based on self-attention weight correction for acoustic and text features. IEEE Access. 2022;10:115732–115743. doi: 10.1109/ACCESS.2022.3219094.
4. Yu C, Su X, Qian Z. Multi-stage audio-visual fusion for dysarthric speech recognition with pre-trained models. IEEE Transactions on Neural Systems and Rehabilitation Engineering. 2023;31:1912–1921. doi: 10.1109/TNSRE.2023.3262001.
5. Aouani H, Ayed YB. Speech emotion recognition with deep learning. Procedia Computer Science. 2020;176:251–260. doi: 10.1016/j.procs.2020.08.027.
6. Kumbhar HS, Bhandari SU. Speech emotion recognition using MFCC features and LSTM network. In: 2019 International Conference on Computing, Communication, and Intelligent Systems (ICCCIS). 2019. doi: 10.1109/ICCCIS48478.2019.8974491.
7. Wang J, Xue M, Culhane R, Diao E, Ding J, Tarokh V. Speech emotion recognition with dual-sequence LSTM architecture. In: 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2020. doi: 10.1109/ICASSP40776.2020.9054629.
8. Yin T, Dong F, Chen C, Ouyang C, Wang Z, Yang Y. A spiking LSTM accelerator for automatic speech recognition application based on FPGA. Electronics. 2024;13(5):827. doi: 10.3390/electronics13050827.
9. Kadiri SR, Gangamohan P, Gangashetty SV, Alku P, Yegnanarayana B. Excitation features of speech for emotion recognition using neutral speech as reference. Circuits, Systems, and Signal Processing. 2020;39(9):4459–4481. doi: 10.1007/s00034-020-01377-y.
10. Zhang S, Zhao X, Tian Q. Spontaneous speech emotion recognition using multiscale deep convolutional LSTM. IEEE Transactions on Affective Computing. 2019;1–1. doi: 10.1109/TAFFC.2019.2947464.
11. Tariq Z, Shah SK, Lee Y. Speech emotion detection using IoT-based deep learning for health care. In: 2019 IEEE International Conference on Big Data (Big Data). 2019. doi: 10.1109/BigData47090.2019.9005638.
12. Abdelhamid AA, *et al*. Robust speech emotion recognition using CNN+LSTM based on stochastic fractal search optimization algorithm. IEEE Access. 2022;1–1. doi: 10.1109/ACCESS.2022.3172954.
13. Kerkeni L, Serrestou Y, Mbarki M, Raoof K, Mahjoub MA, Cleder C. Automatic speech emotion recognition using machine learning. In: Social Media and Machine Learning. 2019. doi: 10.5772/intechopen.84856.
14. Mustaqeem, Sajjad M, Kwon S. Clustering-based speech emotion recognition by incorporating learned features and deep BiLSTM. IEEE Access. 2020;8:79861–79875. doi: 10.1109/ACCESS.2020.2990405.
15. Abdul Qayyum AB, Arefeen A, Shahnaz C. Convolutional neural network (CNN) based speech-emotion recognition. In: 2019 IEEE International Conference on Signal Processing, Information, Communication & Systems (SPICSCON). 2019.
16. Mesaros A, *et al*. Detection and classification of acoustic scenes and events: Outcome of the DCASE 2016 challenge. IEEE/ACM Transactions on Audio, Speech, and Language Processing. 2018;26(2):379–393. doi: 10.1109/TASLP.2017.2778423.
17. Satt A, Rozenberg S, Hoory R. Efficient emotion recognition from speech using deep learning on spectrograms. Interspeech. 2017;2017:1089–1093. doi: 10.21437/Interspeech.2017-200.
18. Tzirakis P, Zhang J, Schuller B. End-to-end speech emotion recognition using deep neural networks. In: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2018. doi: 10.1109/ICASSP.2018.8462677.
19. Sun TW. End-to-end speech emotion recognition with gender information. IEEE Access. 2020;8:152423–152438. doi: 10.1109/ACCESS.2020.3017462.
20. Zhao Z, *et al*. Exploring deep spectrum representations via attention-based recurrent and convolutional neural networks for speech emotion recognition. IEEE Access. 2019;7:97515–97525. doi: 10.1109/ACCESS.2019.2928625.
21. Ren Y, *et al*. FastSpeech: Fast, robust and controllable text to speech. Advances in Neural Information Processing Systems. 2019;32.
22. Mao Q, Dong M, Huang Z, Zhan Y. Learning salient features for speech emotion recognition using convolutional neural networks. IEEE Transactions on Multimedia. 2014;16(8):2203–2213. doi: 10.1109/TMM.2014.2360798.
23. Yoon S, Byun S, Jung K. Multimodal speech emotion recognition using audio and text. In: 2018 IEEE Spoken Language Technology Workshop (SLT). 2018.
24. Abdelhamid AA, *et al*. Robust speech emotion recognition using CNN+LSTM based on stochastic fractal search optimization algorithm. IEEE Access. 2022;1–1. doi: 10.1109/ACCESS.2022.3172954.
25. Aggarwal A, *et al*. Two-way feature extraction for speech emotion recognition using deep learning. Sensors. 2022;22(6):2378. doi: 10.3390/s22062378.
26. Grondin F, Glass J, Sobieraj I, Plumbley MD. Sound event localization and detection using CRNN on pairs of microphones. arXiv. 2019. 1910.10049.
27. Jin P, Si Z, Wan H, Xiong X. Emotion classification algorithm for audiovisual scenes based on low-frequency signals. Applied Sciences. 2023;13(12):7122. doi: 10.3390/app13127122.
28. Anandappa, Mudnal K. Analysis of emotions through speech recognition. Journal of Scientific Research and Technology. 2024;1(3):30–34. doi: 10.61808/jsrt95.
29. Tariq Z, Shah SK, Lee Y. Speech emotion detection using IoT-based deep learning for health care. In: 2019 IEEE International Conference on Big Data (Big Data). 2019. doi: 10.1109/BigData47090.2019.9005638.
30. Wani TM, Gunawan TS, Qadri SAA, Mansor H, Kartiwi M, Ismail N. Speech emotion recognition using convolution neural networks and deep stride convolutional neural networks. In: 2020 IEEE International Conference on Artificial Intelligence in Engineering and Technology (IICAIET). 2020.
31. Khalil RA, Jones E, Babar MI, Jan T, Zafar MH,

Alhussain T. Speech emotion recognition using deep learning techniques: A review. IEEE Access. 2019;7:117327–117345.
doi: 10.1109/ACCESS.2019.2936124.
32. Ardan F, Ciardi FC, Conci N. Speech emotion recognition and deep learning: An extensive validation using convolutional neural networks. IEEE Access. 2023;11:116638–116649.
doi: 10.1109/ACCESS.2023.3326071.
33. Meng H, Yan T, Yuan F, Wei H. Speech emotion recognition from 3D log-Mel spectrograms with deep learning network. IEEE Access. 2019;7:125868–125881. doi: 10.1109/ACCESS.2019.2938007.
34. Alluhaidan AS, Saidani O, Jahangir R, Nauman MA, Neffati OS. Speech emotion recognition through hybrid features and convolutional neural network. Applied Sciences. 2023;13(8):4750. doi: 10.3390/app13084750.
35. Fayek HM, Lech M, Cavedon L. Towards real-time speech emotion recognition using deep neural networks. In: 2015 International Conference on Signal Processing and Communication Systems (ICSPCS). 2015.
36. Aggarwal A, *et al*. Two-way feature extraction for speech emotion recognition using deep learning. Sensors. 2022;22(6):2378. doi: 10.3390/s22062378.