International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 11 Issue: 04 | Apr 2024 www.irjet.net p-ISSN: 2395-0072

UNVEILING SPEECH EMOTIONAL SPECTRUM THROUGH SOUND USING CONVOLUTIONAL NEURAL NETWORKS

Prof. H. Sheik mohideen[1], A. Aslam Sujath[2], P. Pradeep[3], E. Selvakumar[4]

[1]Assistant Professor, Department of CSE, Government College of Engineering, Srirangam, Tamil Nadu, India
[234]UG student, Department of CSE, Government College of Engineering, Srirangam, Tamil Nadu, India

---

Abstract - *Research in Speech Emotion Recognition (SER) has garnered significant attention, particularly in the realm of Human-Computer Interaction (HCI) with a focus on personal assistants and assistive robots. This method involves analyzing subtle tones and pitches in speech, utilizing aural cues to classify human emotions like calm, happy, sad, anger, fear, surprise, and disgust. Annotated datasets like RAVDESS facilitate this research, containing recordings of actors expressing various emotions. Deep learning techniques, especially convolutional neural networks (CNNs), are emerging as powerful tools for processing emotional speech signals. CNNs automatically learn hierarchical representations from raw data, making them adept at capturing complex patterns in audio signals. This approach enhances human-technology interactions by enabling machines to recognize and respond to human emotions conveyed through language. Thus, the integration of SER into HCI research contributes to improving interactive computer systems' design.*

Keywords: **Speech Emotion Recognition, Preprocessing, CNN Classification, Feature Extraction (MFCC, ZCR, CHROMA)**

1. INTRODUCTION

Human computer intelligence is an upcoming field of research which aims to make computers learn from experiences and decide how to a particular situation. This has resulted in improved interaction between users and the computer. With the help of certain algorithms and procedures, the computer can be made fit to detect various characteristics present in the audio sample and deduce emotion underlying. In the field of human-computer interaction, aims to develop a Speech Emotion Recognition System (SERS) using Convolutional Neural Networks (CNN). The selection of the RAVDESS dataset, which provides a diverse set of emotional audio samples, adds depth to our approach. The main goal is to train the system to classify spoken words into seven different emotions. By using deep learning techniques, particularly CNNs, we aim to equip machines with the ability to decipher human emotions and respond effectively. The importance of this project lies in its potential to improve applications ranging from virtual assistants to emotion recognition technology. As technology becomes more integrated into our daily lives, machines' ability to understand and adapt to human emotions will become increasingly important. Our SERS project aims to bridge this gap and create more intuitive and empathetic interactions between humans and machines. Ultimately, our focus is not just on innovation, but on improving the overall user experience. In navigating the complex landscape of human emotion, our voice emotion recognition system will play a key role in driving more meaningful and responsive interactions in the ever-evolving field of artificial intelligence.

1.1 CNNs in speech emotion recognition

Convolutional neural networks (CNNs) are a valuable tool for speech emotion recognition (SER) due to their ability to effectively analyze the intricate patterns found in audio signals. Originally designed for image processing, CNNs have been successfully adapted to handle continuous data like audio, showcasing impressive performance in extracting both basic and advanced features crucial for emotion recognition. In the realm of SER, CNNs employ hierarchical feature learning to pinpoint variations in pitch, intensity, prosody, and other acoustic attributes that signify different emotional states. Moreover, CNN's parameter sharing approach aids in reducing the number of trainable parameters and enhancing the model's adaptability to diverse emotional expressions.

2. REALTED WORK

In recent years, the application of speech emotion recognition has witnessed widespread adoption in the realm of human-computer interaction, offering machines the ability to comprehend and learn human emotions. However, despite significant advancements, the performance of emotion recognition systems falls short of researchers' expectations. Addressing this, two primary challenges in speech emotion recognition are identified: the identification of effective speech emotion features and the construction of a suitable recognition model. Previous studies have explored various feature parameters to enhance emotional recognition tasks utilized pitch frequency, short-term energy, formant frequency, and chaotic characteristics, constructing a 144-dimensional emotion feature vector achieved encouraging results by combining energy, zero crossing rate, and first-order derivative parameters for speech emotion recognition. Despite the progress, the challenge of high dimensionality and feature redundancy persists, necessitating the filtration

of characteristic parameters with higher distinguishability. The Fisher criterion, a classical linear decision method, proves effective in feature selection, as demonstrated by Huang et al., resulting in a notable increase in emotion recognition. To further enhance speech emotion recognition, an effective emotion recognition model is crucial. Various classifiers, including Gaussian mixture model (GMM), artificial neural network (ANN), and support vector machine (SVM), have been employed, with SVM standing out for its advantages in solving complex recognition problems. In an effort to optimize SVM classifiers, researchers have explored innovative algorithms such as the leaping algorithm and integrated systems combining hidden Markov models (HMM) and SVM. In this study, a speech emotion recognition method is proposed based on a decision tree SVM model with Fisher feature selection. Following feature extraction, Support Vector Machine (SVM), Adaptive Boosting (AdaBoost), and Gaussian Mixture Model (GMM) are the commonly used classifiers. However, most of these approaches are unable to characterize the temporal variation of the emotion expression in speech signals. This approach establishes a high- performance decision tree SVM classifier by calculating the degree of emotional confusion, enabling a two-step classification process involving rough and fine classification. By applying Fisher criterion, redundant features are filtered out to obtain an optimal feature set, contributing to enhanced emotional classification performance. The contributions of this paper encompass the adoption of the Fisher criterion to improve emotion recognition performance, the proposal of an algorithm for dynamic decision tree structure determination, and the integration of Fisher criterion with decision tree SVM, further optimized through genetic algorithms. These efforts are implemented and evaluated on the CASIA Chinese speech emotion corpus and the EMO-DB Berlin speech corpus, showcasing the potential for improved emotion recognition rates.

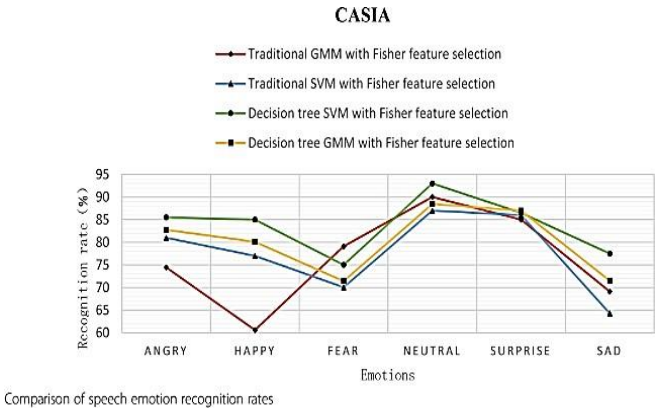

Fig-2.1: Comparison of speech emotion recognition rates using CASIA

Table -1: EMO-DB audio samples

Emotions	Count of audio samples
Anger	127
Bored	81
Disgust	46
Anxiety	69
Happy	71
Sad	62
Neutral	79

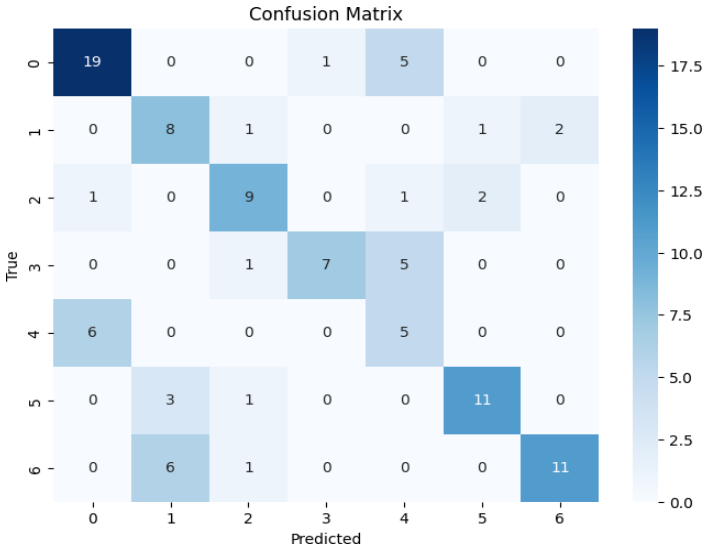

Fig-2.3: Confusion matrix of SVM model with EMO-DB

3. PROPOSED METHODOLOGY

3.1 Methodology

The proposed system for speech emotion recognition (SER) is a combination of the RAVDESS dataset and a convolutional neural network (CNN) architecture. The system effectively integrates Mel frequency cepstral coefficients (MFCC), zero crossing rate (ZCR), and chroma features to classify emotions. The RAVDESS dataset provides a diverse set of labeled emotional speech samples that are vital for training and evaluating SER systems. The MFCC extracts a compressed representation of spectral characteristics, the ZCR captures temporal dynamics, and the chroma function encodes tonal content. These features are concatenated into the CNN's input feature vector, making both spectral and temporal information available. The CNN architecture comprises convolutional, pooling, and fully connected layers that are trained on labeled datasets using supervised learning techniques. Techniques such as dropout regularization and early stopping prevent overfitting during

training. The model's performance on different test sets is evaluated using metrics such as accuracy and the F1 score. Fine-tuning and optimization strategies further refine the model, resulting in a system that can be deployed in real-world applications. This integrated approach provides a robust solution for SER systems with potential applications in the fields of human- computer interaction and emotional computing. The proposed system has the potential to be a game-changer in speech emotion recognition, as it combines different techniques to classify emotions, making it more efficient and effective.

3.2 System Architecture

The architecture for speech emotion recognition consists of two distinct phases: training and testing shown in Fig 3.1.1 During the training phase, machine learning models are trained using labelled datasets, and audio samples are pre-processed and feature-extracted using techniques such as MFCC, ZCR, and chroma features. Supervised learning techniques such as convolutional neural networks (CNN) are utilized to optimize model parameters with algorithms such as gradient descent. Performance evaluation is carried out using a validation dataset. During the testing phase, the trained model is used in real-time to analyse incoming audio streams in a real application or system. The input audio samples undergo preprocessing and feature extraction, similar to the training phase. The extracted features are then used to make predictions regarding the emotional content of speech. These predictions can inform decisions, provide feedback, and trigger actions within your application or system.

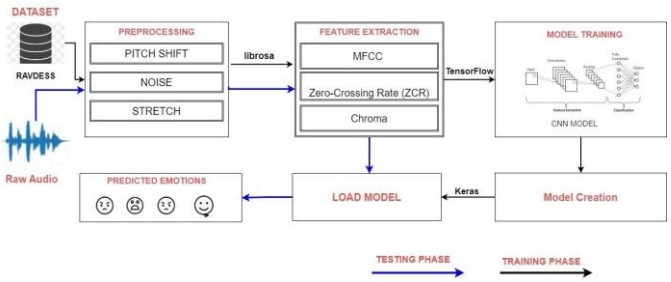

Fig-3.2.1: System Architecture3.2 Methodology

3.3 RAVDESS Dataset

In advancing speech emotion recognition (SER), this system focuses on leveraging the potential of the Ryerson Audio-Visual Database of Emotional Speech (RAVDESS). This part of RAVDESS contains 1440 files: 60 trials per actor x 24 actors = 1440. RAVDESS is made up of 24 professional actors (12 women, 12 men) who pronounce his two lexically matching statements in Neutral North American accents. Verbal emotions include expressions of happy, sad, anger, fear, surprise, neutral and disgust. Each facial expression is generated with two emotional intensities (usually strong), and an additional neutral facial expression is added. The multimodal nature of RAVDESS with different emotional expressions by different actors adds complexity to our approach and train a robust convolutional neural network (CNN) model for speech emotion recognition (SER). This inherent diversity allows our model to capture a wide range of emotional nuances, contributing to the effectiveness of understanding and classifying emotions in language. It aims to improve the model's generalization ability across different emotional contexts and ultimately improve its performance in real-world applications.

Table -2: RAVDESS audio samples

Emotions	Count of audio samples
Anger	192
Happy	192
Disgust	192
Angry	192
Fear	192
Sad	192
Neutral	288

3.4 Data Acquisition and Annotation

In the process of collecting data for speech emotion recognition (SER), it was decided to use the Ryerson Audio-Visual Database of Emotional Speech (RAVDESS). The RAVDESS dataset serves as a primary source of raw audio data and provides a rich collection of diverse emotional expressions. This dataset consists of recordings from multiple actors, contributing variations in voice pitch, accent, and speaking style. The decision to use the RAVDESS dataset is consistent with the goal of training a robust his SER model, as it covers a wide range of emotional states such as happiness, sadness, anger, surprise, fear, and neutral expressions. The multimodal nature of datasets containing audio samples makes the approach more complex and allows the model to capture the nuances present in different audio modalities. Ethical considerations are taken into account throughout the data collection process and when obtaining consent from those contributing to the dataset. Privacy and confidentiality will be maintained in accordance with ethical guidelines. By selecting the RAVDESS dataset, each audio file is labeled with the corresponding emotional state to ensure consistent annotation. This labeled dataset serves as the basis for training and evaluating machine learning models. The accessibility and well-documented nature of the RAVDESS dataset streamlines the data collection process and enables effective integration into SER systems. Annotation involves assigning labels to the collected audio data that indicate the emotional state expressed in each recording.

Emotion labels may include categories such as happiness, sadness, anger, surprise, fear, and neutral. To ensure the reliability of the dataset, it is important to maintain consistency in assigning emotion labels.

3.5 Data preprocessing

Data augmentation is an important step in using the RAVDESS dataset to improve the diversity and robustness of speech emotion recognition (SER) datasets. Using the librosa library implementation, it applies various transformations to the audio data, such as pitch shifting, time stretching, and adding noise, to improve the generalization of the model. Different acoustic conditions and emotional expressions can be simulated, allowing the model to better generalize to unseen instances.

3.5.1 Pitch shifting

Pitch shifting is a fundamental technique in speech emotion recognition (SER) that allows the frequency characteristics of an audio signal to be changed while preserving its temporal characteristics. In SER, pitch shifting serves as a data augmentation method to increase the diversity of the training data set. The SER model detects pitch variations by applying a pitch shift to the original speech sample, allowing it to better generalize to different speech features and emotional expressions shown in Fig.3.5.1. This technique adjusts the audio signal's frequency up or down without affecting its duration. Pitch shifting implemented in Python through signal processing libraries such as Librosa allows researchers to simulate changes in vocal tone, contributing to a more comprehensive and robust training dataset for the SER model.

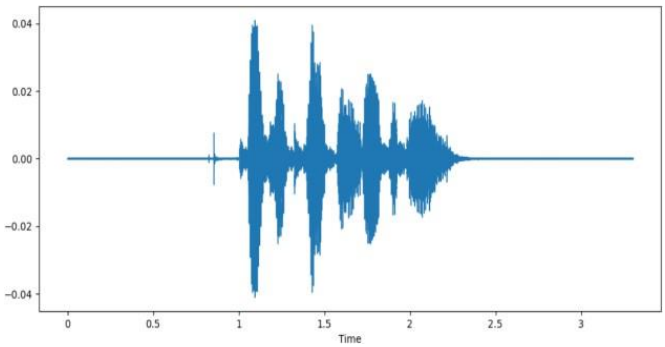

Fig-3.5.1: Pitch Shifting

3.5.2 Time stretching

Time stretching facilitates the modification of the duration of speech signals while preserving their pitch or frequency content. Within SER, time stretching serves as a crucial data augmentation method, enriching the training dataset's variability. By adjusting speech samples' duration, time stretching allows SER models to learn from a broader spectrum of temporal patterns and speech rhythms, ultimately enhancing their capacity to identify emotional expressions across diverse speech speeds and styles shown in Fig.3.5.2. This technique is typically implemented through signal processing libraries like Librosa in Python.

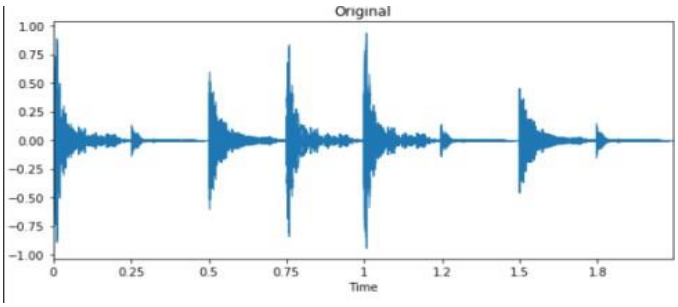

Fig-3.5.2: Time stretching

3.5.3 Adding Noise

Adding noise is a common technique in audio signal processing that introduces random fluctuations to an audio signal. This acts as a data augmentation method that increases the robustness of the training dataset. By injecting random noise into audio samples, researchers can simulate real-world environmental conditions and acoustic distortions, making SER models more robust to noise in real-world applications. This technique superimposes random noise, typically generated from a Gaussian distribution, onto the original audio signal. The added noise level can be controlled to achieve the desired amplification level. SER models gain accuracy and generalization ability by adding noise and learning how to distinguish emotional data from background and other noise.

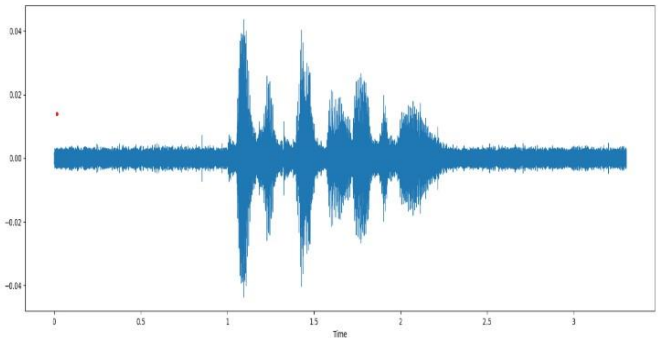

Fig-3.5.3: Adding noise

3.6 Feature Extraction

A variety of Python libraries, including librosa, Numpy, Pandas, and Matplotlib, are used for feature extraction from audio recordings because of their user-friendly interface and extensive capabilities. For feature extraction from audio samples, load the audio file using

librosa and extract various features such as Mel Frequency Cepstral Coefficient (MFCC), Zero Crossing Rate (ZCR), Chroma, Root Mean Square Energy (RMSE).

3.6.1 Mel-frequency Cepstral coefficients (MFCC)

The MFCC represents the short-term power spectrum of a sound, captures the spectral envelope of an audio signal, and is often used as a feature in speech and audio processing tasks for capturing the acoustic signal shown in Fig 3.6.1.

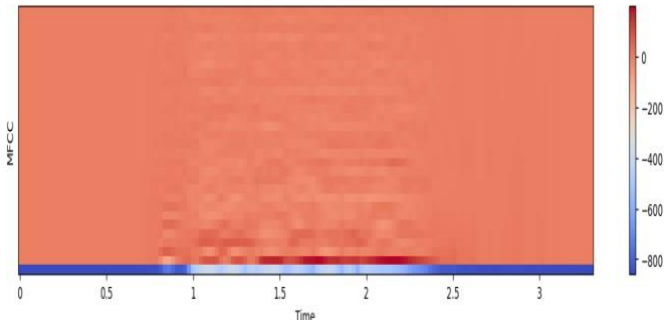

Fig-3.6.1: Mel-frequency Cepstral coefficients

3.6.2 Zero Cross Rate (ZCR)

ZCR measures how quickly the sign of an audio signal changes and provides information about the frequency content and periodicity of the signal. This is useful for tasks such as speech onset recognition and rhythmic analysis.

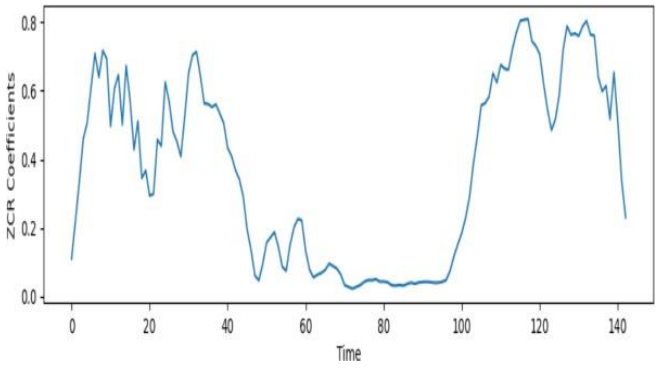

Fig-3.6.2: Zero Cross Rate

3.6.2 Chroma

The chroma function describes the distribution of pitch classes in an audio signal that is invariant to changes in timbre or octave, and is useful for tasks involving analysis of harmonic content, such as musical genre classification and chord recognition shown in Fig 3.6.3.

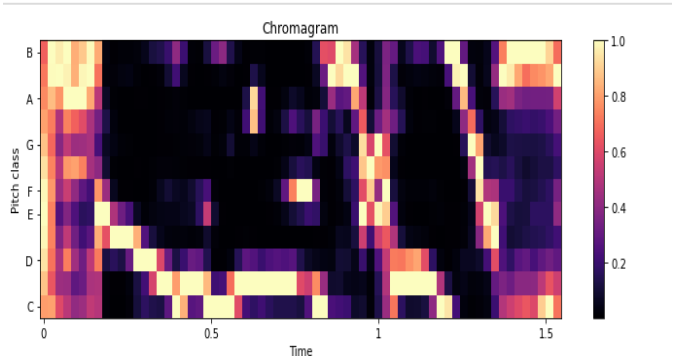

Fig-3.6.3 Chroma

The extracted features were converted into a Pandas Data Frame in tabular format to facilitate further analysis.

	0	1	2	3	4	5	6	7	8	9	...	2257	2258	2259	2260	2261
0	0.028809	0.042969	0.057129	0.052734	0.065430	0.062988	0.064941	0.063965	0.049316	0.044922	...	8.507217	8.082827	8.828318	8.973294	2.93084
1	0.046387	0.061523	0.076660	0.054688	0.068359	0.065918	0.069824	0.077637	0.087402	0.132812	...	8.507217	8.082827	8.828318	8.973294	2.93084
2	0.030273	0.043457	0.056152	0.057129	0.062500	0.062012	0.065918	0.066895	0.070801	0.069336	...	8.507217	8.082827	8.828318	8.973294	2.93084
3	0.035156	0.048340	0.061035	0.056152	0.061523	0.061035	0.064941	0.067871	0.078613	0.087891	...	8.507217	8.082827	8.828318	8.973294	2.93084
4	0.023926	0.036133	0.048805	0.053711	0.060547	0.066895	0.077637	0.087891	0.092285	0.089844	...	8.507217	8.082827	8.828318	8.973294	2.93084
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
11195	0.016113	0.044922	0.064453	0.074707	0.081055	0.061035	0.050781	0.036133	0.031250	0.033691	...	8.507217	8.082827	8.828318	8.973294	2.93084
11196	0.020996	0.029297	0.038086	0.029785	0.032715	0.032715	0.035156	0.035156	0.031738	0.029785	...	8.507217	8.082827	8.828318	8.973294	2.93084
11197	0.025879	0.034180	0.043945	0.032715	0.034668	0.043457	0.046875	0.046875	0.042480	0.031738	...	8.507217	8.082827	8.828318	8.973294	2.93084
11198	0.022461	0.028809	0.035156	0.034180	0.041016	0.060547	0.075195	0.069336	0.059570	0.040527	...	8.507217	8.082827	8.828318	8.973294	2.93084
11199	0.026367	0.032715	0.040039	0.032227	0.043945	0.069336	0.095703	0.091797	0.078125	0.053223	...	8.507217	8.082827	8.828318	8.973294	2.93084

11200 rows × 2267 columns

3.7 CNN Model

Convolutional neural network (CNN) architectures tailored for audio emotion recognition are highly integrated with different components, each of which plays a critical role in accurately classifying emotions from audio signals. At the core of a CNN, an input layer is used to receive a concatenated feature vector extracted from an audio sample, such as Mel-frequency cepstral coefficients (MFCC), zero-crossing rate (ZCR), and chroma features. These feature vectors provide a comprehensive representation of the spectral and temporal speech characteristics of the audio signal. Successive convolutional layers form the backbone of the CNN and perform feature extraction by convolving the input feature map with a learnable filter. These filters capture different patterns and structures in the input data and help the model detect subtle differences in emotional expressions. Activation functions such as Rectified Linear Units (Relu) introduce nonlinearity to the model, allowing it to learn complex mappings between input features and emotion classes. The pooling layer plays an important role in down sampling the feature map created by the convolutional layer, reducing overfitting and improving computational efficiency by preserving the most salient features while reducing the spatial dimension. The fully connected layer

serves as the final stage of the CNN. The extracted features are integrated and classified into emotion classes. This layer uses the hierarchical representation learned from previous layers to make informed predictions. Regularization techniques such as dropout and batch normalization prevent overfitting during training and improve the model's generalization performance on unseen data. Optimization algorithms such as stochastic gradient descent (SGD) using Adam and Momentum efficiently update model parameters to minimize the loss function, allowing CNNs to accurately classify emotions based on audio signals. The output layer uses SoftMax activations to compute the probability distribution between emotion classes and provides a measure of the model's confidence in its predictions shown in Fig 3.7. This comprehensive CNN architecture combines spectral and temporal features with deep learning techniques to produce robust speech emotion recognition with potential applications in human-computer interaction, affective computing, and mental health monitoring. It offers the way for more empathetic and intelligent systems.

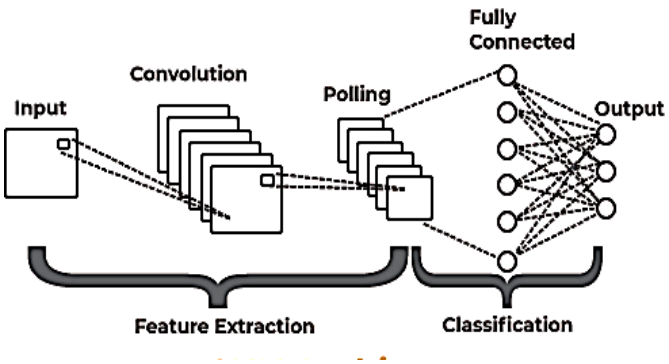

Fig-3.7: CNN Model Architecture

Table -3: Neural Network Training Parameters

Parameter	Description
Optimizer algorithm	Adam optimizer
Learning rate	0.001
Batch size	32
Number of epochs	50 to 100
Dropout	0.5 or 50%
Loss function	Categorical cross entropy
Activation Functions (Hidden Layer)	Relu
Activation functions (Output Layer)	Softmax

The Adam optimizer has been employed to find the optimal node weights and minimize prediction error. Adam is derived from adaptive moment estimation. The Adam optimizer modifies the learning rate throughout the training process with its many advantages, the Adam optimizer is highly recommended as a default optimizer technique and servers as a useful benchmark for deep learning projects. Compared to other optimization algorithms, it works more quickly, uses less memory, and needs less running. The Adam optimizer combination of gradient descent with momentum and the RMSP algorithm.

In the output layer, the SoftMax activation function is used, which is used for multiclass classification. The output of a SoftMax is a vector containing probabilities for each class. The vector's probability for all potential outcomes or classes sum to one.

Categorical Cross-Entropy loss has been used in the study, Cross-Entropy, often known as loss, is a performance metric for classification models whose output is a probability value between 0 and 1. As the predicted probability gets closer 1, log loss drops gradually. Cross-entropy loss grows then predicted probability deviates from the actual label, and as a result, errors are penalized

4. EXPERIMENTAL EVALUATION

4.1 Environment Specification

It represents the experimental setup of the study. The research work takes place on an AMD-Ryzen 3 CPU. Furthermore, the machine has 8GB of RAM and an integrated graphics card. All of the models are built in Python and run on deep learning libraries like Keras and TensorFlow.

Table -4: Experimental Setup

Process name	S.No	Action
Input	1.	Collection of input audio samples include 7 differentemotions.
Environment Configuration	2.	Anaconda, Jupyter Notebook
	3.	Import all necessary libraries and packages
Directories Configuration	4.	Load the audio samples
	5.	Load the directories for training, testing and create validation on training data
Training and Testing	6	Build a CNN models
	7	Fine-tuned models by adding the Relu and SoftMax activation function

Model Compilation	8	The model complies with Adam optimizer and a learning rate 0.001
	9	Set 50 epochs for model training
	10	As model checkpoint, use the validation loss to monitor
	11	Save model
Performance Report	12	Generate classification report
	13	Generate model accuracy and loss reports
Prediction	14	Load the trained model
	15	Predict the type of speech emotions.

4.2 Evaluation metrics

The effectiveness of the statistical, ML, or DL model is assessed using evaluation metrics. To evaluate a study's proposed model, it is crucial to use a variety of evaluation metrics. Metrics for evaluation are crucial for ensuring models' performance. Model predictive or classification efficiency can be measured by accuracy, precision, recall, and the f1-score are measured for all emotions show in Table 4.2.

4.2.1 Accuracy

Accuracy is the percentage of images correctly predicted from all the predictions. The following Eq. (9) describes how the accuracy is stated:

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

The accuracy score measures the number of correct predictions (TP + TN) made by a model in relation to the total number of predictions (TP + TN + FP + FN) made. The abbreviations for "true positive", "true negative", "false positive", and "false negative" are "TP", "TN", "FP", and "FN" respectively.

Accuracy of the CNN model = 95.75%

4.2.2 Precision

The precision, which measures the percentage of truly right positive outcomes, is determined by using the formulae.

$$\text{Precision} = \frac{TP}{TP+FP}$$

4.2.3 Recall

By comparing the number of true positive findings to the total number of actual positive samples (TP + FN), the recall value is used to gauge the accuracy of positive predictions. The following equation is used to determine the recall value:

$$\text{Recall or TPR} = \frac{TP}{TP+FN}$$

4.2.4 F1-Score

To measure how well a model performs, researchers utilize measures like the F1-score, which is calculated by taking the harmonic mean of the model's precision and recall. It is defined as Eq:

$$\text{F1 – Score} = 2 \times \frac{Recall \times Precision}{Recall + Precision}$$

Table -5: Evaluation Metrics

	Precision	Recall	F1-Score
Angry	0.96	0.94	0.95
Disgust	0.96	0.94	0.95
Fear	0.97	0.93	0.95
Happy	0.92	0.96	0.94
Neutral	0.95	0.99	0.97
Surprise	0.98	0.97	0.97
Sad	0.98	0.95	0.97

4.3 Result analysis and deployment

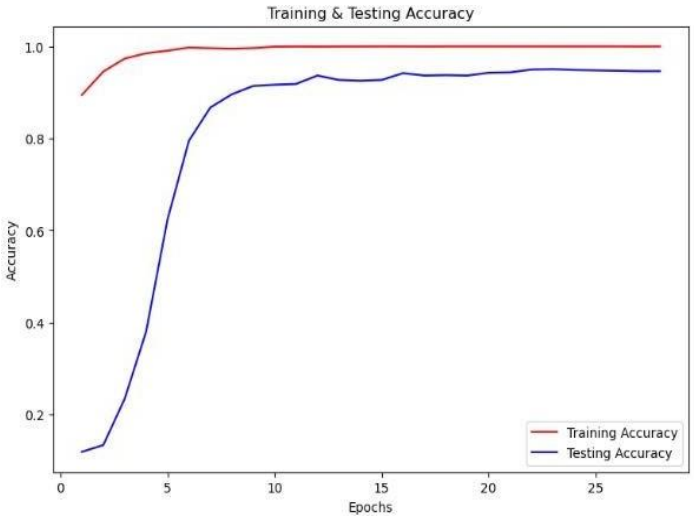

Fig-4.3.1: Training & Testing accuracy graph

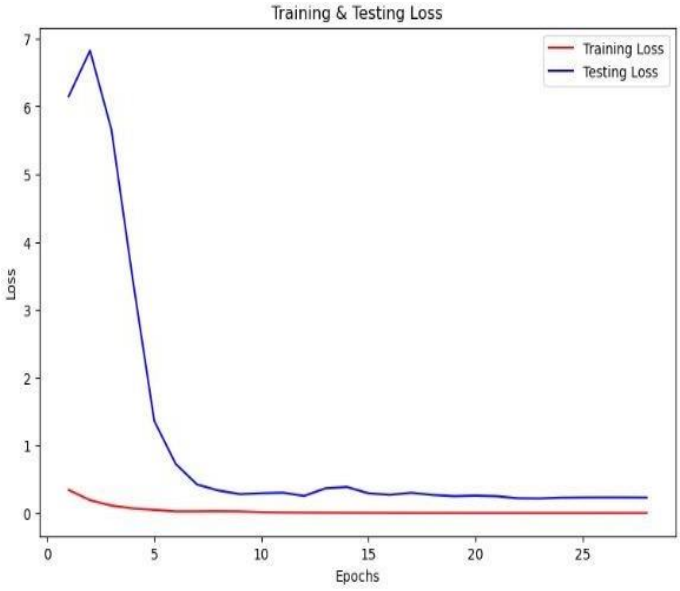

Fig-4.3.2: Training & Testing loss graph

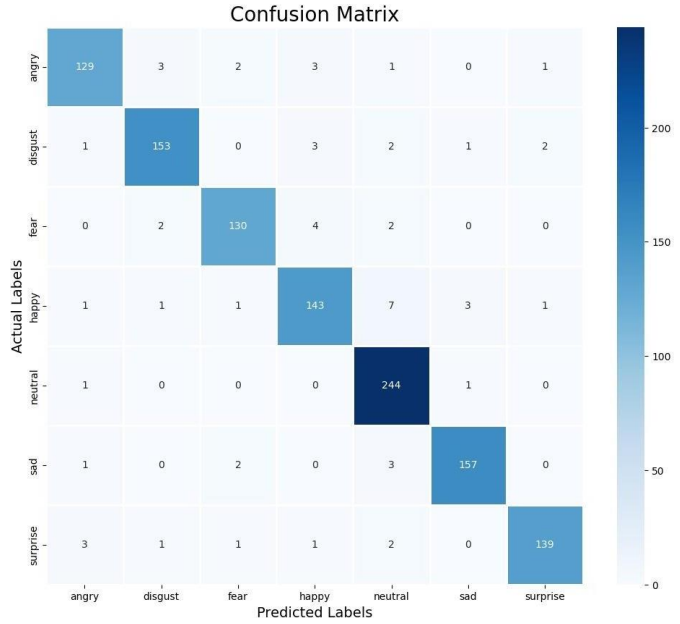

Fig-4.3.3:Confusion matrix

5.CONCLUSION

Speech emotion recognition (SER) is at the forefront of human-computer interaction, providing deep insight into the complexity of human emotions conveyed through speech based on audio feature analysis and machine learning algorithms, the system crosses disciplinary boundaries and has applications in a variety of fields, including healthcare, education, and entertainment. SER is focused on exploring the RAVDESS dataset, a large collection of emotional speech samples spanning the full range of human experience and speech expression. The SER journey begins with the extraction and analysis of key audio features. This task can be accessed through powerful Python libraries such as librosa, NumPy, Pandas, and Matplotlib. These libraries allow researchers to harness the potential of audio data and gain insight into the spectral and temporal properties of emotional speech. Among the myriad features extracted, the Mel frequency cepstral coefficient (MFCC) emerges as a fundamental element that captures the subtle patterns of vocal colour and articulation underlying emotional expression. In addition to MFCC, features such as zero crossing rate (ZCR), chroma, and root mean square energy (RMSE) provide additional analytical dimensions and reveal rhythmic, tonal, and dynamic aspects of emotional speech. Beyond academia, SER offers practical applications in a variety of real-world scenarios, each driven by the promise of improving human interaction and understanding. In the healthcare field, SER has the potential to revolutionize the diagnosis and treatment of mental illness by enabling emotion-aware virtual assistants to provide compassionate support to those in need. Educational institutions benefit from SER-driven personalized learning experiences, where emotion-aware tutoring systems adapt their teaching approaches based on students' emotional state and engagement. Additionally, in customer service, SER enables sentiment analysis of call centre interactions, allowing companies to provide a more empathetic and personalized service experience. The entertainment and media industries are harnessing the power of SER to create immersive experiences that engage viewers on a deeper emotional level. From emotion-aware content recommendation systems to interactive gaming experiences that adapt to player emotions in real-time, SER drives innovation in storytelling and audience engagement. Additionally, SER provides valuable insight into consumer sentiment and emotional responses in market research and advertising, guiding the development of targeted marketing campaigns and product strategies. Looking ahead, the future of SER is full of possibilities thanks to advances in machine learning, natural language processing, and affective computing. Researchers continue to push the boundaries of SER technology, ushering in a new era of emotionally intelligent systems that will enrich our lives and further deepen our understanding of human emotions. Through collaborative efforts and a commitment to ethical and inclusive design practices, SER fosters more caring and meaningful interactions between people and technology, ultimately creating a more caring and connected world.

6.FUTURE WORK

Speech emotion recognition (SER) includes several important directions, including consideration of multimodal approaches that integrate audio, video, and text data to improve emotion recognition accuracy. Deep learning architectures such as recurrent neural networks (RNNs) and transformer models promise improved SER performance. Transfer learning and domain adaptation techniques

facilitate the generalization of the SER model to new domains or languages with limited labelled data. Real-time emotion recognition systems, contextual emotion recognition models, and advances in emotion understanding and generation are also important areas of focus. Additionally, consideration of ethical and social implications such as privacy, bias, and fairness are critical for the responsible development and use of his SER technology in a variety of applications. Through these research directions, the future of SER aims to advance the state of the art and enable more empathetic and emotionally intelligent human-computer interactions.

REFERENCES

[1] A. Milton, S. Sharmy Roy, S. Tamil Selvi, "SVM Scheme for Speech Emotion Recognition using MFCC Feature," International Journal of Computer Applications, vol. 69, no. 9, May 2013.

[2] Akshath Kumar B.H, Nagaraja N Poojary, Dr. Shivakumar G S, "Speech Emotion Recognition Using MLP Classifier," International Journal of Scientific Research in Computer Science, Engineering and Information Technology, vol. 7, no. 4, 2021.

[3] Fu Wang, Linhui Sun, Sheng Fu, "Decision tree SVM model with Fisher feature selection for speech emotion recognition," EURASIP Journal on Audio, Speech, and Music Processing, 2019.

[4] G. Liu, W. He and B. Jin, "Feature Fusion of Speech Emotion Recognition Based on Deep Learning," International Conference on Network Infrastructure and Digital Content (IC-NIDC), pp. 193-197, Guiyang, China, August 22–24, 2018.

[5] H. O. Nasereddin and A. R. Omari, "Classification Techniques for Automatic Speech Recognition (ASR) Algorithms used with Real Time Speech Translation," Computing Conference, London, UK, pp. 200-207, 2017.

[6] K. Tarunika, R. B. Pradeeba and P. Aruna, "Applying Machine Learning Techniques for Speech Emotion Recognition," 9th International Conference On Computing, Communication and Networking Technologies.

[7] Masato Akagi, Reda Elbarougy, "Feature Selection Method or Real-time Speech Emotion Recognition," International Committee for Coordination and Standardization of Speech Databases and Assessment Technique, November 2017.

[8] S. Prasomphan," Improvement of Speech Emotion Recognition with Neural Network Classifier by Using Speech Spectrogram", 2015, International Conference on Systems, Signals and Image Processing (IWSSIP), London, UK, 10-12 September, pp. 72–76, 2015.

[9] T. Özseven, "Investigation of the Effect of Spectrogram Images and Different Texture Analysis Methods on Speech Emotion Recognition," Applied Acoustics, pp. 70–77, 2018.

[10] Ye Sim Ülgen Sonmez, Asaf Varol, "New Trends in Speech Emotion Recognition," Institute Of Electrical And Electronics Engineers, June 2019.