ISSN: 2583-7346
DOI: https://doi-ds.org/doilink/06.2023-34882542/0004062023, June 2023, Volume-1, Issue-3, pp. 14-24
International Journal of Advances in Engineering Architecture Science and Technology

Design & Development of Emotion Recognition for Speech Data

Badugu Srinivasu[1], Bellamkonda Meghana[2]

[1]Professor, Department of Information Technology, Stanley College of Engineering & Technology for Women, Abids, Hyderabad, India.

[2] UG Student, Department of Information Technology, Stanley College of Engineering & Technology for Women, Abids, Hyderabad, India.

Received: 08-06-2023 **Accepted:** 09-06-2023 **Published:** 09-06-2023

Abstract

Background: Raw emotions have an impact on communication in real life. Human emotion can be instantly recognised by the way you move, the way you smile, the way you talk, postures, and words.
Objectives: The key goal of speech-emotion recognition (SER) is to determine the emotional characteristics of speech, regardless of semantic content. The SER system merely includes a collection of approaches that classify and process speech signals to identify feelings that are present in them. Deep learning (DL) techniques are used to comprehend what is possible for people, including their ability to perceive emotions in communication.
Methods / Statistical Analysis: The RAVDESS (Ryerson Audio-Visual Database of Emotional Speech and Song) dataset served as the basis for our study on detecting emotions by means of speech using CNNs in deep learning. Each audio sample is 4-5 seconds long and was recorded by 24 actors (12 men and 12 women) in one of eight diverse emotional states: calm, happy, sad, angry, fearful, surprise, disgust and neutral. DL models extract high-level characteristics from unprocessed audio inputs, making emotion recognition more precise and reliable.
Findings: Relevant qualities are often taken out of the raw speech signal as an integral part of feature extraction step of the DL models used for SER. The proposed method classified the speaker's emotional states with a general accuracy of 62%.
Applications / Improvements: It is possible to enhance human-machine interaction, improve speech-based customer service systems, and create human-centred assistive technologies with disabilities by recognising feelings from speech signals.

Key words: Deep learning, neural networks, machine learning, speech recognition; emotion recognition.

1. Introduction

Deep learning-based speech emotion recognition is the method of identifying and computing a speaker's emotional state as determined by the words they use. The use of artificial neural networks in deep learning, a branch of machine learning, to model and find solutions to complicated issues can be recognised. Speech and emotion detection have multiple kinds of applications, including speech-based human-computer interaction and healthcare.

* Bellamkonda Meghana, UG Student, Department of Information Technology, Stanley College of Engineering & Technology for Women, Abids, Hyderabad, India.
Email: bellamkondameghana999@gmail.com

ISSN: 2583-7346
DOI: https://doi-ds.org/doilink/06.2023-34882542/0004062023, June 2023, Volume-1, Issue-3, pp. 14-24
International Journal of Advances in Engineering Architecture Science and Technology

The structure of this research is as follows: Background information on speech recognition, the emotion recognition system, emotion recognition applications, and some further appropriate information are included in Section 2. Explaining the literature review in Section 3 A detailed suggested approach is discussed in Section 4. A complete experimental setup is shown in Section 5 for categorising voice signals based on emotion. In Section 6, a detailed review of the results and analysis is discussed. In Section 7, a conclusion is presented.

2. Background

The efficiency of deep learning algorithms for recognising speech emotions depends on a number of factors, including the quality of the voice signal, the amount and sort of data used for training, the model's learning architecture, and the chosen hyperparameters.

Emotion recognition

The process of identifying emotions using technology to locate and analyse human emotions is done based on various cues such as facial expressions, speech patterns, and physiological signals. Emotional recognition is the establishment of algorithms and models that accurately detect and classify emotions expressed by humans. Emotional recognition uses various methods and techniques, including computer vision, speech processing, and physiological signal analysis. Machine learning and artificial intelligence systems that extract information from emotions are used to classify input data.

Speech recognition

Using technology, speech recognition transforms spoken words into typed text. The procedure of recognising and transcribing spoken real-time language is commonly referred to as speech recognition. Systems that detect speech typically involve three main components: the acoustic model, the language model, and the decoder. Considering the speech context, the language model predicts the most likely words or phrases. The decoder combines the speech output and language models to generate the final transcription.
Speech recognition has a wide range of practical applications, including voice-controlled assistants, automated transcription, and dictation software. In addition to language learning, it can be effective for disability support and assistive technology.

Speech Emotion Recognition

The approach to identifying and analysing the emotional content of spoken language using technology is called speech emotion recognition (SER). SER creates algorithms and models that precisely identify and categorise speech-expressed emotions.
SER typically entails assessing several speech acoustic characteristics, including pitch, intensity, and spectral traits. Both artificial intelligence and machine learning techniques are frequently used to train models that identify patterns in the features and categorise them in accordance with certain emotions.
There are numerous possible uses for speech and emotion detection. It can be used to monitor and spot mental health problems such as depression, anxiety, and bipolar disorder. It can also be useful to comprehend how consumers' emotional reactions to goods and promotional efforts develop more effective marketing strategies.

ISSN: 2583-7346
DOI: https://doi-ds.org/doilink/06.2023-34882542/0004062023, June 2023, Volume-1, Issue-3, pp. 14-24
International Journal of Advances in Engineering Architecture Science and Technology

CNN

Characteristics are taken from the CNN (convolutional neural network) input image by convolving it with a collection of filters or kernels. Then, for classification or prediction, the features are sent via several layers of neurons.
To increase their precision and generalisation, CNNs are combined with other deep learning methods, including transfer learning, data augmentation, and regularisation capabilities.

Deep Learning

Artificial neural networks are used in deep learning, a branch of machine learning, to gather large datasets and produce predictions. To automatically recognise intricate patterns and relationships in data, several layers of deep learning algorithms are interconnected by neurons. Deep learning models are used in applications including recommender systems, natural language processing, and audio and image recognition.

Advantages

- Improved accuracy: Models for deep learning can extract and learn complex representations of speech features, allowing for more accurate emotion recognition than traditional approaches to machine learning.
- Robustness: Deep learning models handle noisy and varied data, which is essential for emotion recognition tasks where speech signals can vary widely in regard to accent, tone, and background noise.
- Flexibility: Models can be trained to recognise different emotions and be easily adapted to various tasks or contexts.
- Feature extraction: Automatic extraction of relevant features from the raw speech signal, reducing manual feature engineering.
- Scalability: Large datasets are handled and can be easily scaled to larger datasets, allowing for more accurate recognition of a wider range of speech samples.

Disadvantages

- Data requirements: Restricted supply of labelled datasets; voice and emotion detection tasks can be difficult as deep learning models require a lot of labelled data to be trained efficiently.
- Computational requirements: Training necessitates a lot of processing power, which can be found with high-performance computing systems or specialised hardware like GPUs or TPUs.
- Interpretability: Models can be tricky to read, which makes it difficult to comprehend how the model creates predictions and restricts diagnostic errors or bias.
- Generalisation: If the training data is biased or confined in scope, models may not generalise successfully to new or unexplored datasets.
- Overfitting: Overfitting of training data would lead to poor performance on new data.

Applications

Speech and emotion recognition can be utilised in this application to improve interaction between humans and computers by allowing computers to recognise and react properly to the

ISSN: 2583-7346
DOI: https://doi-ds.org/doilink/06.2023-34882542/0004062023, June 2023, Volume-1, Issue-3, pp. 14-24
International Journal of Advances in Engineering Architecture Science and Technology

emotional state of the person using them. The user experience could be enhanced by, for instance, having virtual assistants like Siri or Alexa react differently.
• In the medical field, speech and emotion recognition is used to track a patient's emotional state and spot symptoms of sadness or anxiety, which are frequently accompanied by changes in speech patterns. Patients who struggle with communication or speech impairments can also use it to help themselves express what they are feeling.
• Marketing and advertising: Speech and emotion recognition can be used to examine social media postings, reviews, and feedback from customers about their satisfaction, mood, and emotional engagement.
• Education: Teachers can change the way they teach to increase student engagement and learning outcomes by using speech and emotion recognition to track students' emotional states during lectures or classes.

3. Literature Survey

This project attempted to use the inception net for solving the emotion recognition problem; various databases have been explored, and the IEMOCAP database is used as the dataset for carrying out my experiment. I trained my model using TensorFlow. An accuracy rate of about 38% is achieved. In the future, real-time emotion recognition [5] can be developed using the same architecture.
The problem arises when speech samples are of different sizes; for this type of data, the input feature matrix may be mostly sparse. Though standard feed-forward MLP is a powerful tool for classification problems, an extremely sparse matrix may not yield favourable results; however, experiments with a properly tuned [8] MLP network should be interesting. In this paper, we attempt to solve the problem of SER using a feature learning scheme based on deep convolutional neural networks. The speech signal is represented as spectrograms, which act as the input to deep CNNs. The CNN model, consisting of three convolutional and three fully connected layers, extracts features from these spectrograms and outputs predictions for the seven emotion classes. [10].

4. Methodology

This section explains the proposed methodology and the emotion database used for research.

Emotion Database

In this paper, we present a study on the design and development of speech and emotion recognition using CNNs in deep learning, with the RAVDESS (Ryerson Audio-Visual Database of Emotional Speech and Song) as our primary dataset. The dataset consists of 7356 audio files, each 4-5 seconds long, recorded by 24 actors (12 male and 12 female) in seven emotional states: neutral, calm, happy, sad, angry, fearful, and disgusted.

Architecture

Deep learning-based SER involves data pre-processing, feature extraction, model training, and evaluation. Data pre-processing consists of cleaning and formatting the audio data, while feature extraction requires transforming the audio data into a format that is used as input for

ISSN: 2583-7346
DOI: https://doi-ds.org/doilink/06.2023-34882542/0004062023, June 2023, Volume-1, Issue-3, pp. 14-24
International Journal of Advances in Engineering Architecture Science and Technology

the deep learning model. Model training involves using labelled data to train the deep learning model, while evaluation entails testing the model on newly generated, unseen data.

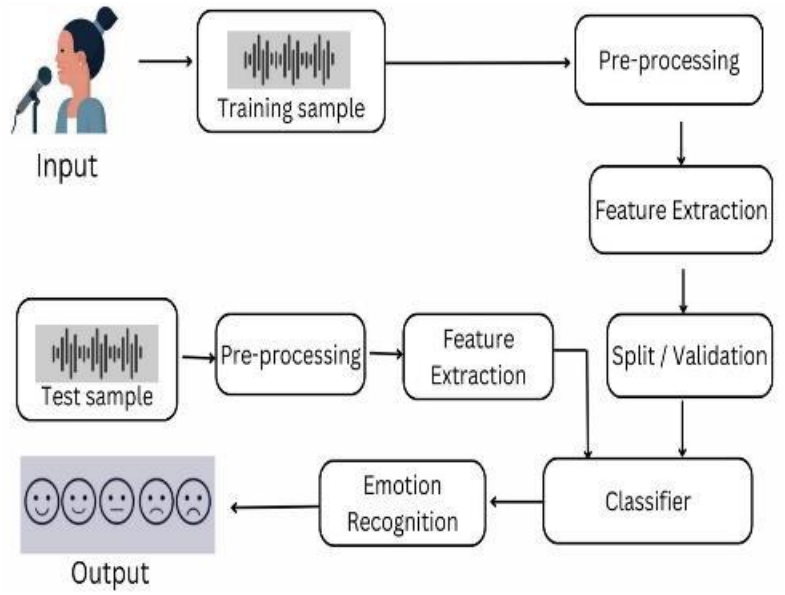

Figure 1. System Architecture

Data pre-processing techniques play a crucial role in improving the accuracy of the data given by using deep learning and convolutional neural networks (CNN). Data pre-processing techniques are used to extract significant features from speech signals that can be fed into CNN for training and classification.

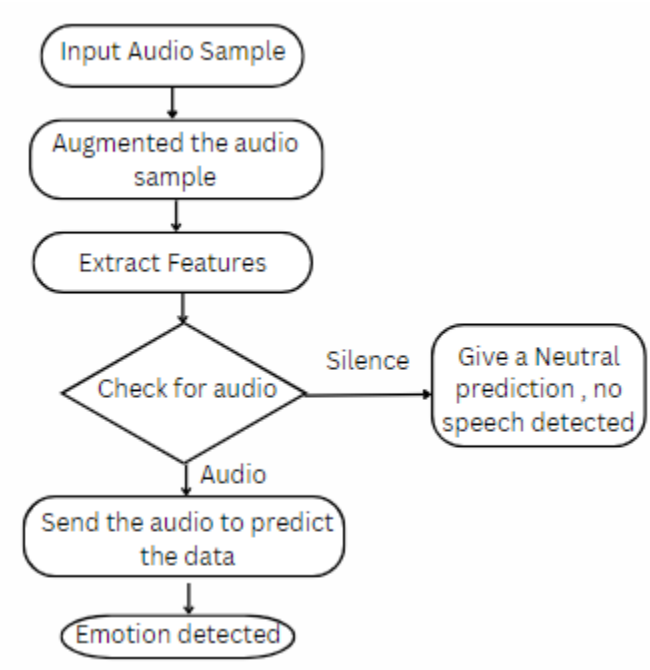

Figure 2. Training Flow

ISSN: 2583-7346
DOI: https://doi-ds.org/doilink/06.2023-34882542/0004062023, June 2023, Volume-1, Issue-3, pp. 14-24
International Journal of Advances in Engineering Architecture Science and Technology

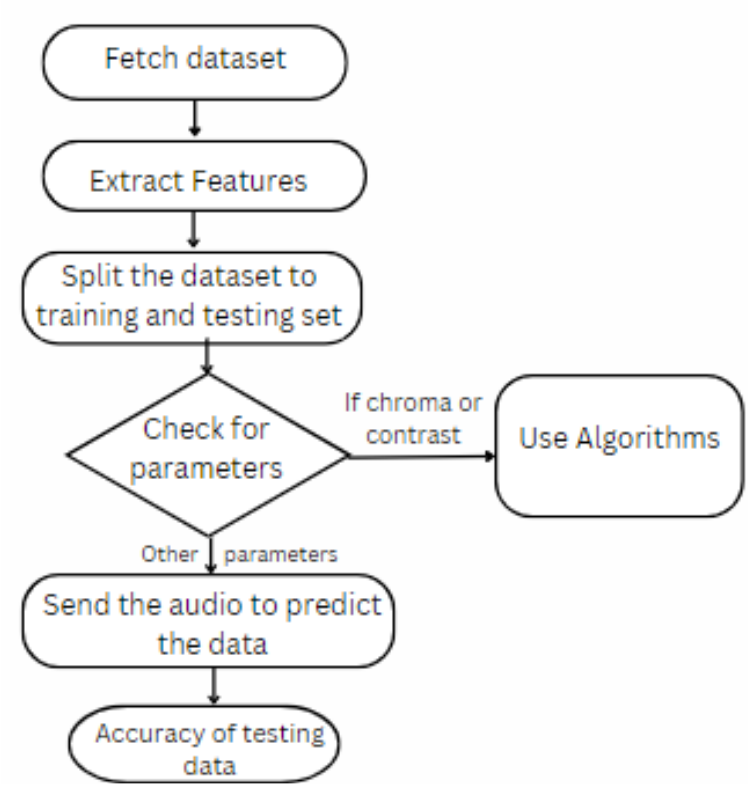

Figure 3. Testing Flow

Module Description

Acoustic pre-processing, such as denoising and segmentation, is the method used to establish meaningful units for this signal. Feature extraction is utilised to identify rare event features available in the signal, and lastly, classifiers map extracted feature vectors to relevant emotions. The first stage of speech-based signal processing, speech enhancement, is carried out where noisy components are removed. In the second stage, two parts are involved: feature extraction and feature selection.

5. Experimental Setup

This part provides an explanation of the experimental setup and the deep learning packages used in emotion recognition.

Libraries Used

A Python package for analysing music and audio is called Librosa. The Python package NumPy is used to manipulate arrays, and it provides functions for working on matrices, Fourier transforms, and the field of linear algebra. A cross-platform package for graphic charting and data visualisation, Matplotlib. TensorFlow is an open-source toolkit for numerical computation that is compatible with Python. A Python interface is provided by the open-source software package known as Keras. The TensorFlow library interface is provided by Keras and is designed to quickly experiment with deep neural networks. Data is loaded and stored in a pickle.

ISSN: 2583-7346
DOI: https://doi-ds.org/doilink/06.2023-34882542/0004062023, June 2023, Volume-1, Issue-3, pp. 14-24
International Journal of Advances in Engineering Architecture Science and Technology

Workflow

- Audio Signal: The audio signal is the raw data captured from the microphone or read from a file.
- Pre-processing Techniques: The audio signal is pre-processed to remove noise, normalise the volume, and apply other techniques to boost the standards of the audio data.

- Feature Extraction: The pre-processed audio signal is transformed into a set of features used as input for the CNN. This typically involves applying techniques like Mel Frequency Cepstral Coefficients (MFCCs) to extract information about the audio signal.
- Convolutional Neural Network: The extracted features from the audio signal are fed into a CNN, which learns to recognise flows in the input data that are associated with different emotions.
- Classification: The output of the CNN is a probability distribution over different emotion category. A classification algorithm (e.g., Softmax) to determine the most likely emotion category for the input signal.
- Emotion Category: The final output is the predicted emotion category for the given audio signal (e.g., happy, sad).

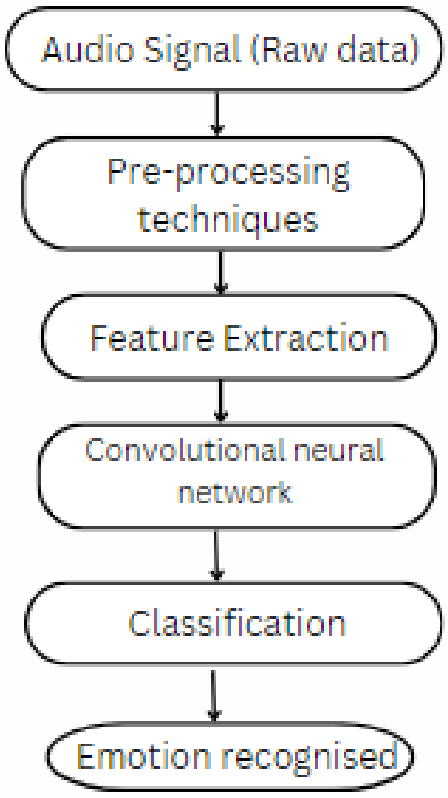

Figure 4. Workflow

6. Result & Analysis

The data model used to predict emotions has an accuracy rate of roughly 62%.

ISSN: 2583-7346
DOI: https://doi-ds.org/doilink/06.2023-34882542/0004062023, June 2023, Volume-1, Issue-3, pp. 14-24
International Journal of Advances in Engineering Architecture Science and Technology

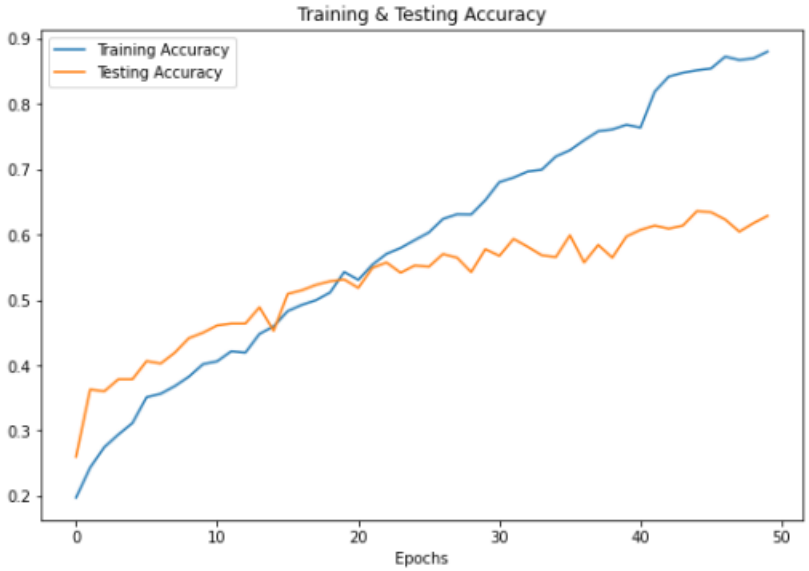

Figure 5. Training & Testing Accuracy

The training dataset is used to train the model, while the test dataset is used to evaluate the performance of the model on unseen data. To plot the accuracy graph for the train and test datasets, we can plot two separate curves on the graph. The x-axis of the graph represents the number of training epochs, while the y-axis gives the accuracy of the model. Ideally, both the train and test accuracy curves should increase steadily without plateauing or decreasing.

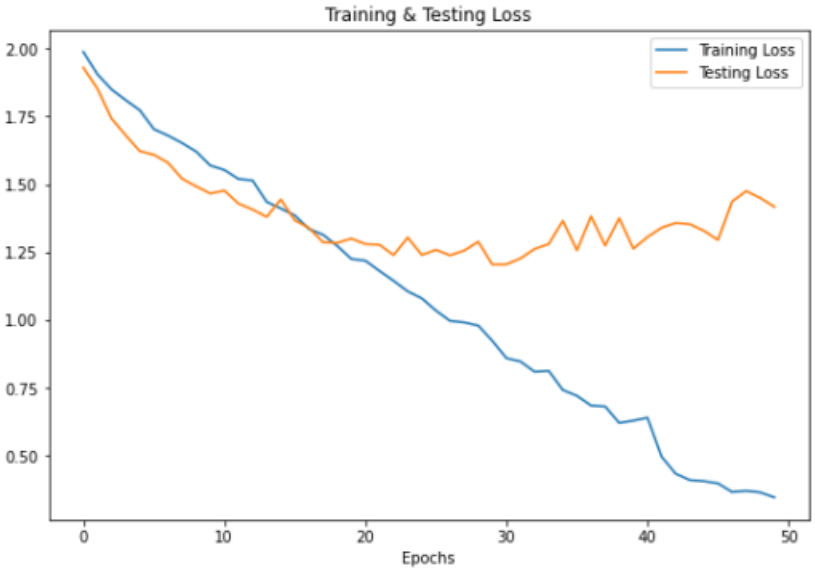

Figure 6. Training & Testing Loss

The model may be overfitting the training data if the training accuracy is higher than the test accuracy. This indicates that the model cannot generalise adequately to new data.
The model may be underfitting if both the train and test accuracy are low. This indicates that the model requires further training because it is unable to recognise the underlying flow of the data.

ISSN: 2583-7346
DOI: https://doi-ds.org/doilink/06.2023-34882542/0004062023, June 2023, Volume-1, Issue-3, pp. 14-24
International Journal of Advances in Engineering Architecture Science and Technology

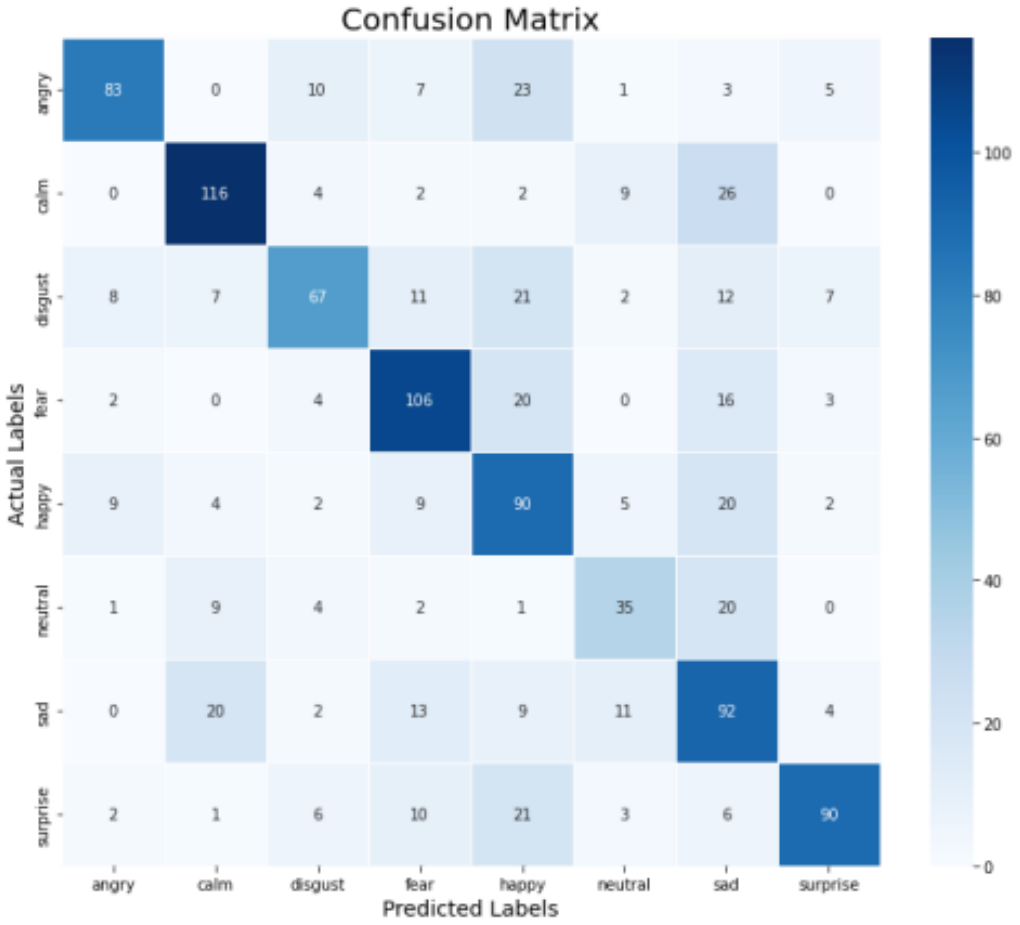

Figure 7. Confusion Matrix

When a model learns to match training data too well, it is overfitted, which causes it to perform poorly on new, untried data. We can employ strategies like early halting or regularisation to avoid overfitting.

Ideally, the confusion matrix should have a high number of correct predictions along the diagonal and a low number of incorrect predictions in the off-diagonal cells. This would indicate that the model is performing well and is accurately recognising the emotion signals.

In conclusion, these investigations show that CNNs can be highly accurate at detecting speech emotions on the RAVDESS dataset. But it's important to remember that the model's performance may differ based on the particular architecture and hyperparameters employed, as well as the particular emotions being categorised.

7. Conclusion

In this study, we used deep learning to analyse certain speech samples. We first loaded the datasets, and then using the wave display and spectrogram functions from the Librosa library, we visualised the various human emotions. The MFCC approach was then used to gather sound features from the samples. We then constructed the model, trained it, and used the Matplotlib library to graphically represent the data. The model's average accuracy after extensive testing using various values is 62%.

ISSN: 2583-7346
DOI: https://doi-ds.org/doilink/06.2023-34882542/0004062023, June 2023, Volume-1, Issue-3, pp. 14-24
International Journal of Advances in Engineering Architecture Science and Technology

Table 1. Detailed Accuracy Report

	Precision	Recall	f1-score	Support
Angry	0.79	0.63	0.7	132
Calm	0.74	0.73	0.73	159
Disgust	0.68	0.5	0.57	135
Fear	0.66	0.7	0.68	151
Happy	0.48	0.64	0.55	141
Neutral	0.53	0.49	0.51	72
Sad	0.47	0.61	0.53	151
Surprise	0.81	0.65	0.72	139
Accuracy			0.63	1080
Macro Average	0.65	0.62	0.62	1080
Weighted Average	0.65	0.63	0.63	1080

The RAVDESS dataset is used as the main dataset in this research to present a study on speech and emotion identification using CNNs in deep learning. The proposed method demonstrated the viability of CNNs for voice and emotion recognition using the RAVDESS dataset.
Future research can concentrate on enhancing the developed method's accuracy by integrating more sophisticated deep learning methods like recurrent neural networks (RNNs) and attention processes. The suggested approach can also be used with other SER datasets to test its generalisation potential.

References

1. Speech Emotion Recognition Using Deep Learning Techniques: A review by Ruhul Amin Khalil, Edward Jones, Mohammad Inayatullah Babar, Tariqullah Jan, Mohammad Haseeb Zafar, And Thamer Alhussain (2019).
2. Speech And Emotion Recognition Using Deep Learning by Dr Senthil Kumar.M, Surendra P, Subhash S (2022).
3. Machine Learning Based Speech Emotions Recognition System by Dr Yogesh Kumar, and Dr Manish Mahajan (2019).
4. Speech Recognition with Deep Learning by Lokesh Khurana, Arun Chauhan, Dr Mohd Naved, Prabhishek Singh (2020).
5. Speech Emotion Recognition Using Deep Learning by Nithya Roopa S., Prabhakaran M, Betty. P (2018).
6. A Comprehensive Review of Speech Emotion Recognition Systems by Taiba Majid Wani, Teddy Surya Gunawan, Syed Asif Ahmad Qadri, Mira Kartiwi and Eliathamby Ambikairajah (2021).
7. Speech Emotion Recognition Using Support Vector Machines by Thapanee Seehapoch, Sartra Wongthanavasu (2013).
8. A Review on Emotion Recognition using Speech by Saikat Basu, Jaybrata Chakraborty, Arnab Bag and Md. Aftabuddin (2017).
9. Speech emotion recognition with acoustic and lexical features by Qin Jin, Chengxin Li, Shizhe Chen, and Huimin Wu (2015).

ISSN: 2583-7346
DOI: https://doi-ds.org/doilink/06.2023-34882542/0004062023, June 2023, Volume-1, Issue-3, pp. 14-24
International Journal of Advances in Engineering Architecture Science and Technology

10. Speech Emotion Recognition from Spectrograms with Deep Convolutional Neural Network by Abdul Malik Badshah, Jamil Ahmad, Nasir Rahim, Sung Wook Baik (2017).