Speech Emotion Recognition using CNN

Akash Sahani[1], Mohit Sancheti[2], Shivam Singh[3], Solunke Akash[4], Mrs. Preeti Satao [5]
[1234] *Student, Computer Engineering, MCT's Rajiv Gandhi Institute of Technology, Mumbai University*
Assistant Professor, MCT's Rajiv Gandhi Institute of Technology, Mumbai University

***Abstract*— Speech Emotion Recognition (SER) is an advanced technology that automatically obtains the emotional countries of speakers by assaying their speech signals. It's a collection of colourful methodologies and algorithms that process and classify speech signals to descry feelings bedded in them. SER takes speech as the carrier of emotion to study the conformation and change of colourful feelings in speech. This enables computers to dissect the speaker's specific emotional situation through speech, making mortal- computer commerce more humanized and intuitive. To enhance the delicacy of an intelligent SER system, a speech emotion recognition model grounded on the point representation of a Convolutional Neural Network (CNN) can be used. CNNs are a type of neural network that can effectively dissect and classify image data, making them well- suited for use in SER systems that reuse speech signals as image-suchlike representations.**

I. INTRODUCTION

Voice is a common and natural way for humans to communicate and convey emotional information. In human-computer interaction, it is important to not only obtain the information of voice signals but also the emotional state in each other's voice signals. This can help improve communication and reduce discomfort in human-computer interaction. However, emotional information is unevenly distributed on speech signals, which makes extracting effective features more challenging and puts forward more strict requirements for network structure. Voice is the most basic and direct emotional carrier in interpersonal communication, conveying both semantic information and the emotional state of the speaker. In addition to voice, there are many other emotional expressions in interpersonal communication, including facial expression and body posture. These non-verbal cues can provide additional context and information about the speaker's emotional state and can be used in conjunction with voice signals to improve emotion recognition.

II. PROBLEM STATEMENT

In the modern era, when developing machine learning models for speech emotion recognition (SER), two primary concerns are typically addressed: reducing the cost complexity and improving the stateof-the-art performance. To tackle these challenges, a novel lightweight convolutional neural network (CNN) model has been proposed that uses plain rectangular kernels and a modified pooling strategy. The model is designed to focus on the frequency features in speech spectrograms, which are used to recognize the hidden emotional features in the input audio. By analyzing these deep frequency features, the model can effectively classify the emotional state of the speaker. Overall, this approach aims to strike a balance between accuracy and efficiency, making it a promising solution for real-world applications of SER

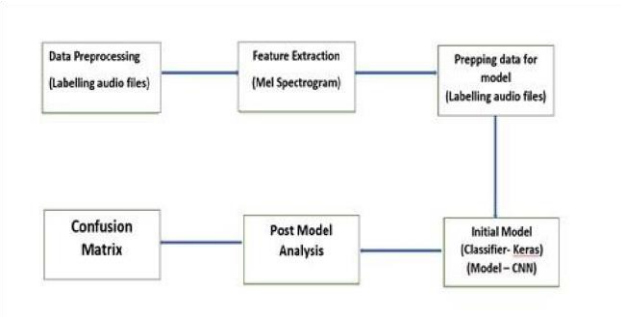
III. LITERATURE REVIEW

Being work in the area of emotion recognition reveals that utmost of the present work relies on verbal analysis for classifying feelings into three orders Angry, Happy, and Neutral. One fashion involves calculating the maximum cross-correlation between the separate time sequences of audio signals. The loftiest degree of correlation between the testing audio train and the training audio train is used as an integral parameter for relating a particular emotion type. Another fashion involves rooting discriminative features with a Boxy SVM classifier for feting Angry, Happy, and Neutral emotion parts only. These ways aim to ameliorate the delicacy of emotion recognition in speech signals.

A study was conducted to compare the performance of emotion bracket using two different machine literacy ways a Support Vector Machine (SVM) and a Multi-Layer Perceptron (MLP) Neural Network. The classifiers used prosodic and voice quality features uprooted from the Berlin Emotional Database using PRAAT tools. The WEKA tool was used for bracket.

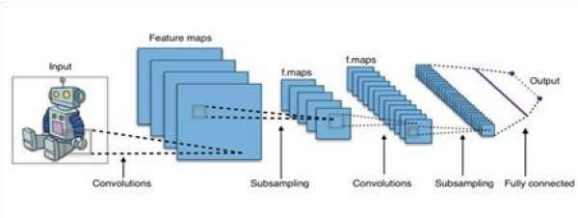
Different parameters were set up for both SVM and MLP to gain optimized emotion bracket. The results showed that while the MLP had better overall emotion bracket performance than the SVM, with an delicacy of 78.69 compared to 76.82 for SVM, the training time for SVM was important faster than for MLP. This suggests that while MLP may give more accurate results, SVM may be a more effective choice in terms of training time.

IV. SYSTEM ARCHITECTURE



V. METHODOLOGY

CNN - Convolutional Neural Network



Convolutional neural networks (CNNs) are highly effective deep learning models that are widely used in various applications. The CNN model is composed of three essential building blocks: convolution layers, pooling layers, and fully connected layers. The input's feature map is generated using the convolution operation, and down-sampling of the feature map is performed using the pooling strategy. The fully connected layers receive the extracted features and transform them into more reliable forms for the final predictions. This process allows the CNN model to analyse complex patterns and structures in the input data and make accurate predictions based on the learned features.

Emotion recognition involves many methods to determine the type of emotion being expressed, but the accuracy of the algorithm determines the output. In some cases, the algorithm may predict the probabilities of different emotions equally, making it difficult to decide the correct emotion. To improve the accuracy and correctly classify the emotion, Convolutional Neural Network (CNN) can be used. CNNs can extract features from raw data and effectively identify complex patterns in the input, making them ideal for emotion recognition tasks. By using a CNN model for emotion recognition, we can achieve higher accuracy and better classification of emotions in the input data.

1.Convolutional layers are used to identify varied length utterances, highlight sections with gaps, and explain feature map sequences.

2.Activation Layer: The functions of the nonlinear activation layer are frequently used for the output of the convolutional layer. Rel Us in our work.

3.Maximum clustering layer: This layer activates the options with the maximum value for the dense layer. This keeps variable length entries in fixed-size arrays of entities.

4.Dense layer

The process of implementing the CNN model- Speech represented as a three-layer image. When using a CNN, consider the first and second derivatives of the speech image in time and frequency. CNN can predict, analyse voice data, CNN can learn from speech and recognize words or utterances.

Using a Convolutional Neural Network (CNN) model for emotion recognition has several advantages.

- First, the prediction time is faster compared to other models, making it suitable for real-time applications.
- Second, the CNN architecture supports parallelization, allowing for the use of multiple processors or GPUs to improve computation time.
- Finally, CNNs have been shown to achieve higher accuracy in classification tasks compared to other machine learning models, making them a popular choice for emotion recognition. Overall, these advantages make CNNs a powerful tool for emotion recognition tasks in various applications.

VI. RESULTS AND DISCUSSION

Home

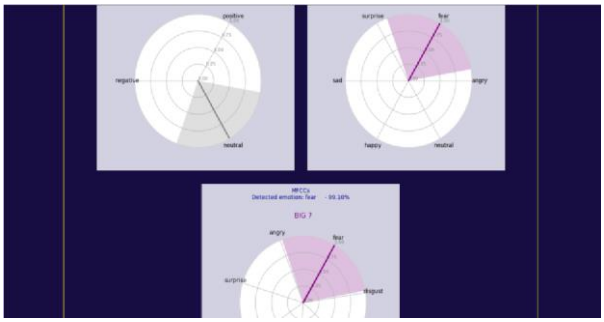
This is the introductory page of the website. This page gives the basic idea about the content of the website. The user can access the About us page for more information about the webpage. The gradient used in the background is added using CSS.

Upload Or Record

This is the upload or record. A user can upload an existing audio file/pre-recorded audio file which contains speech for which he/she wants to recognize emotion for. A user can also record a real time speech for 5 seconds using the record feature. Once the audio file is selected or record is selected user then can recognize emotion using those files.

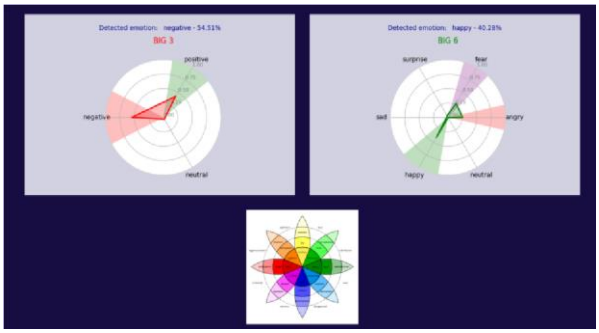
Result 1- Output

This is the output screen which provides all the details which tells about the human gender as well as about their emotion. It contains of two methods. First is MFCCS and second is MEL-Spectrum.



Result 2- MFCCS

This page helps the user to recognize human speech emotion using MFCCS. It determines human speech emotion using MFCCS.

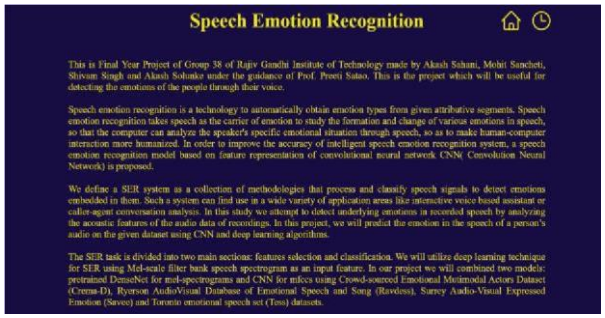


Result 3 – Mel- Spectrum

This page helps the user to recognize human speech emotion using MEL -Spectrum. It determines human speech emotion using Mel- Spectrum.

Recent Activity

This is recent activity page which contains the last 9 recognized results.



About Us

This is the about us page, which contains information about the project.

How to use

This is how to use page, which gives detailed explanation about how to use the project.

VI.CONCLUSION

In conclusion, the implementation of speech emotion recognition using CNN and machine learning algorithms is a significant advancement in the field of natural language processing. This technology not only conveys semantic information but also the emotional state of the speaker. The applications of this technology are vast and can be implemented in various industries. Through the use of deep learning techniques like recurrent neural networks and attention processes, voice emotion recognition has considerably increased in accuracy and effectiveness over time. The future of speech emotion recognition looks promising, and its implementation can lead to a better understanding of human communication and behaviour.

ACKNOWLEDGEMENT

We wish to express our sincere gratitude to Dr Sanjay U. Bokade, Principal and Prof. S. P. Khachane. H.O.D. of Department Computer Engineering of Rajiv Gandhi Institute of Technology for providing us an opportunity to do our project work on “Speech Emotion Recognition”. This project bears on imprint of many peoples. We sincerely thank our project guide Preeti Satao for her guidance and encouragement in carrying out this synopsis work. Finally, we would like to thank our colleagues and friends who helped us in completing project work successfully.

REFERENCE

[1] Kumari S, Perinban D, Balaji M, Gopinath D, Hariharan S -Speech Emotion Recognition Using Machine Learning
[2] Mohammad Rabiei and Alessandro Gasparetto - Recognition of Emotions Based on Speech Analysis, for Applications to HumanRobot Interaction
[3] Nagaraja N Poojary, Dr. Shivakumar G S, Akshath Kumar B.H Speech Emotion Recognition Using SVM Classifier
[4] Vamshidhar Singh -SPEECH BASED EMOTION RECOGNITION USING VOICE.
[5] S. Lalitha; Abhishek Madhavan; Bharath Bhushan; Srinivas Saketh - Speech emotion recognition