IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Recognition Of Emotion Through Speech Using Deep Convolution And LSTM

[1]Dr. T. Praveen Blessington, [2]Devendra Hadke, [3]Snehal Chavan, [4]Prathamesh Chavan, [5]Prathmesh Ingulkar

[1]Guide, [2,3,4,5]Students

[1,2,3,4,5]Department Of Information Technology

Zeal College of Engineering and Research, Narhe, Pune, India-411041

Abstract: The part of emotions in human internal health is extremely important. Since robotic feelings in real scripts are more challenging to detect than other emotions, emotion recognition in real scripts has attracted a lot of attention in affective computing lately. Motivated by the aid of using the multitudinous issues of colorful lengths of audio spectrograms on emotion identification, this paper proposes a Long Short-Term Memory (LSTM) model for speech emotion recognition. originally, a deep convolutional neural network (CNN) interpretation is used to probe deep member-stage capabilities at the foundation of the created image- such as 3 channels of spectrograms. also, a deep LSTM interpretation is followed on the idea of the set up-out member-stage CNN capabilities to seize the temporal reliance amongst all divided parts in an utterance for utterance-stage emotion recognition. Eventually, special emotion recognition results, entered with the aid of using combining CNN with LSTM at further than one length of member-stage spectrograms, are included with the aid of using the use of a score-stage emulsion strategy.

Keywords **–CNN, LSTM, Emotion and AlexNet.**

I. Introduction

Since existing spontaneous emotions in real situations are more complex and difficult to identify than other emotions, emotion recognition in real situations, such as the wild, has received extensive attention in emotion computing.

This article proposes a multi-scale Deep Convolution Long Short-Term Memory (LSTM) frame for the spontaneous recognition of speech sensation, motivated by the various effects of different lengths of audio spectrograms on sensation identification. At first, a profound convolutional brain organization (CNN) model is utilized to decide out profound portion level choices given the made picture like, three channels of spectrograms.

To capture the temporal dependency between all divided segments in associate auditory communication for utterance-level feeling recognition, a deep LSTM model is then acquired based on the learned segment-level CNN options. At the score level, a fusion strategy was used to integrate several sentiment recognition results that were obtained by combining CNN and LSTM at various spectrogram lengths at the segment level.

As human beings' speech is the most natural thing, thanks to specific thoughts. Emotions are important for remote communication in today's digital world because they play a significant role in interpersonal communication. Emotion detection may be a very difficult task as a result, emotions are subjective. There's no common agreement on a way to live or reason with them. Classifying speech as an emotion is challenging because of its subjective nature. This can be simple to watch since this task can be challenging for humans, in addition to machines. Potential applications for classifying speech to emotion are countless, including however not exclusive to, decision centers, AI assistants, counseling, and exactness tests. During this project, we tend to decide to address these issues. We are going to use CNN and LSTM to classify opposing emotions. We tend to separate the speech by speaker gender to probe the connection between gender and the emotional content of speech. There is a spread of temporal and spectral options that may be extracted from human speech. sustainability, and financial desperation. However, these decisions can have devastating consequences on their families and the economy as a whole. In India, where agriculture and related industries contribute to a significant portion of the Gross Value Added, the wrong crop choice can lead to financial strain and even farmer suicide cases. Therefore, the decision of which crop to grow is a critical one and should be made with careful consideration and informed guidance.

II. RELATED WORKS

Since spontaneous emotions in real scenes are more challenging to identify than other emotions, emotion recognition in real scenes, such as the wild, has recently received much attention in affective computing. A multiscale deep convolutional LSTM framework for spontaneous speech emotion recognition is proposed in this paper, inspired by the various effects of different lengths of audio spectrograms on emotion identification. Based on the created image-like, three channels of spectrograms, a CNN model is initially used to learn deep segment-level features. The learned segment-level CNN features are then used in a deep LSTM model to capture the temporal dependence of all divided segments in an utterance for emotion recognition at the utterance level. Finally, a score-level fusion strategy is used to integrate the various emotion recognition results obtained by combining CNN and LSTM at various lengths of segment-level spectrograms.[1]

In Human-Computer Interaction (HCI), emotion recognition from speech signals is a crucial but challenging component. Many well-known speech analysis and classification techniques have been used to extract emotions from signals in the literature on speech emotion recognition (SER). In SER, deep learning methods have recently been proposed as an alternative to traditional methods. The speech-based emotion recognition applications of Deep Learning are the subject of some recent research, and the paper provides an overview of these techniques. The database used, the emotions extracted, the contributions made to speech emotion recognition, and the limitations associated with it are all covered in the review.[2]

Human mental life is significantly affected by feelings. It is a method for communicating one's perspective or mental state to others. The extraction of the speaker's state from their discourse signal is intended to be implied by the expression "Speech Emotion Recognition" (SER). Any keen framework with restricted computational assets can be prepared to recognize or integrate a couple of all-inclusive emotions, like neutral, anger, happiness, and sadness. In this work, spectral and prosodic features are utilized for speech emotion recognition because they both contain emotional information. The Mel-Frequency Cepstral Coefficients (MFCC) is one of the spectral characteristics.

Emotions can be modeled using prosodic features like fundamental frequency, loudness, pitch, speech intensity, and glottal parameters. Each utterance's potential features were extracted to calculate the connection between speech patterns and emotions. The chosen highlights can be utilized to distinguish pitch, which can then be utilized to characterize orientation. Gender is classified using a Support Vector Machine (SVM) in this work. Given the chosen highlights, the feelings were perceived utilizing the Outspread Premise Capability and the Back Spread Organization. It has been exhibited that the spiral premise capability delivers more exact outcomes for feeling acknowledgment than the backpropagation network does.[3]

Discourse feeling acknowledgment is troublesome because of the emotional hole that exists between low-level highlights and abstract feelings. By coordinating staggered highlight learning and model preparation, Profound Convolutional Brain Organizations (DCNN) have shown striking outcomes in spanning the semantic hole in visual assignments like picture characterization and article discovery. The utilization of a DCNN to connect the close-to-home hole between discourse signals is the subject of this examination. First, we extract three Mel spectrogram channel logs that are comparable to the RGB image representation and serve as the DCNN's input—static, delta, and delta-delta. The AlexNet DCNN model is then used to learn high-level feature representations for each segment that is divided from an utterance after it has been trained on a large ImageNet dataset. A Discriminant Short Lived Pyramid Planning (DTPM) system is used to join the learned segment-level components. DTPM creates a global feature representation at the utterance level by combining optimal Lp-norm pooling and temporal pyramid matching. The subsequent stage is to utilize Straight Help Vector Machines (SVM) to order feelings. Another captivating finding is that the

DCNN model can extricate full of feeling discourse includes genuinely well in spite of having been pre-prepared for picture applications. The recognition performance is significantly improved by further fine-tuning the target emotional speech datasets.[4]

When using voice for SER, the accuracy of the recognition increases as more data are used. In particular, a significant amount of data is required for deep learning. However, when using an existing data set, the length of the data can be inconsistent and the size of the set is limited. The audio files of utterances of varying lengths comprise the dataset used in this study. In this paper, deep learning methods such as a Multi-Layer Perceptron (MLP) and a CNN were used to extract one-dimensional data from speech files and train two-dimensional Mel-spectrogram images. Additionally, audio files were pre-processed and shortened to less than two seconds to increase the test accuracy.[5]

Automatic Speech Emotion Recognition (SER) has received a lot of attention in recent years. The enhancement of the human-machine interface is the primary objective of SER. In lie detectors, they can also be used to monitor a person's psychological and physiological state. Speech emotion recognition has recently found use in forensics and medicine as well. Pitch and prosody features were used to identify seven different emotions in this study. The majority of speech features used in this study are time-domain. Emotions have been classified using a support vector machine (SVM) classifier.[6]

Speech emotion recognition algorithms have been the subject of numerous studies. However, the majority relies on selecting the appropriate speech acoustic features. In this paper, we propose a novel emotion recognition algorithm that combines speaker gender information with speech acoustic features. We want to get the most out of the rich information in raw speech data without using artificial means. Emotion recognition in speech typically necessitates the manual selection of appropriate traditional acoustic features for use as classifier input. The network automatically selects important information from the raw speech signal using deep learning algorithms so that the classification layer can perform emotion recognition. Emotional information that cannot be directly mathematically modelled as a speech acoustic characteristic may be prevented from being lost. To further improve recognition accuracy, we also include gender information for speakers in the proposed algorithm. A gender information block and a Residual Convolutional Neural Network (R-CNN) are combined in the proposed algorithm. These two blocks receive the raw speech data simultaneously. The R-CNN network sorts the speech data into the appropriate emotional categories and extracts the necessary emotional data. Three public databases with various language systems serve as the basis for evaluating the proposed algorithm.[7]

III. EXISTING MODEL

Researchers have been increasingly becoming aware of emotional instability and more and more fake behavior in people so that they can keep the person in front of them happy. So, one way of combating the emotional instability of a person is to visit a psychiatrist and take therapy sessions. But the cost charged by psychiatrists is very so common people are never able to afford the treatment and continue to suffer.

Another approach is to create a model using deep learning with can detect the emotion of a person through audio. But up until now there wasn’t much advancement in the artificial intelligence domain, but due to advancements in recent years, we are to create powerful models with are able to handle and process even as much as data in Terabytes.

IV. PROPOSED MODEL

So, in order to address the above-mentioned drawbacks, we need to create an effective Emotion Detection Model. This model will be able to identify the features present in the audio file, features as pitch, sampling rate, bass, and many more. It will be segregated distinctively and it will be done by using Mel-Spectrogram.

We are converting the audio file into a Mel-Spectrogram to identify the features present as Mel-Spectrogram creates an image just like a heat map but the high-value features will have brighter color thereby making them easy to identify.

Then we are extracting the features which are more readily identifiable in the Mel-Spectrogram and then check for their values so determine which emotion the audio file is trying to express.

The extraction of features from the Mel-Spectrogram is done by using CNN and LSTM Models.

The dataset in this study consists of hundreds of audio files with both male and female speeches with different emotions expressed through the audio.

The methodology for this project focuses on creating a CNN model to filter out the unrequired information from the audio input and using LSTM to identify the emotion expressed through the speech.

CNN:

CNN uses spatial correlations between the input data and itself. Some input neurons are connected between each concurrent layer of the neural network. The hidden neurons are the focus of the local receptive field.

A CNN is a type of network architecture for deep learning algorithms that are used for image recognition and other tasks that require processing pixel data. In deep learning, there are other kinds of neural networks, but CNNs are the preferred network architecture for identifying and recognizing objects.

LSTM:

Long Momentary Memory, or LSTM, is a technique utilized in profound learning. It is a collection of recurrent neural networks (RNNs) that can learn semipermanent dependencies, which can be helpful in sequence prediction problems specifically. With the exception of image-like single data points, LSTM is able to process the entire knowledge sequence thanks to its feedback connections. This has applications in speech recognition, machine interpretation, and other areas. LSTM is a remarkably sensible RNN that performs exceptionally well in a wider range of problems than is typical.

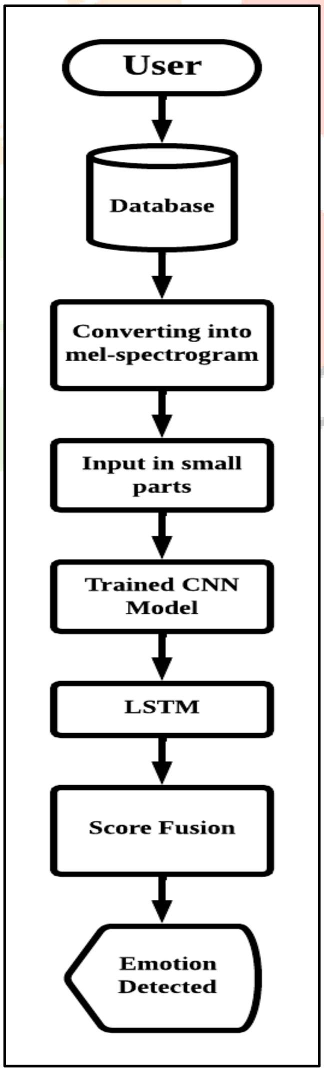


Fig 1 – Proposed model plan

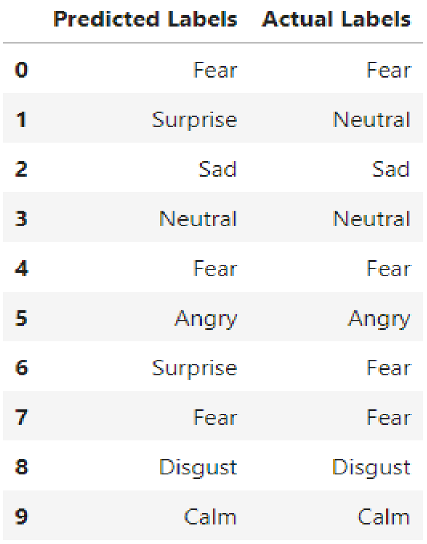
V. RESULTS AND DISCUSSION

The project aimed to develop a system for emotion detection through speech analysis. Emotions play a crucial role in human communication, and being able to accurately detect and understand emotions from speech can have various applications, such as improving human-computer interaction, mental health monitoring, and sentiment analysis in customer service.

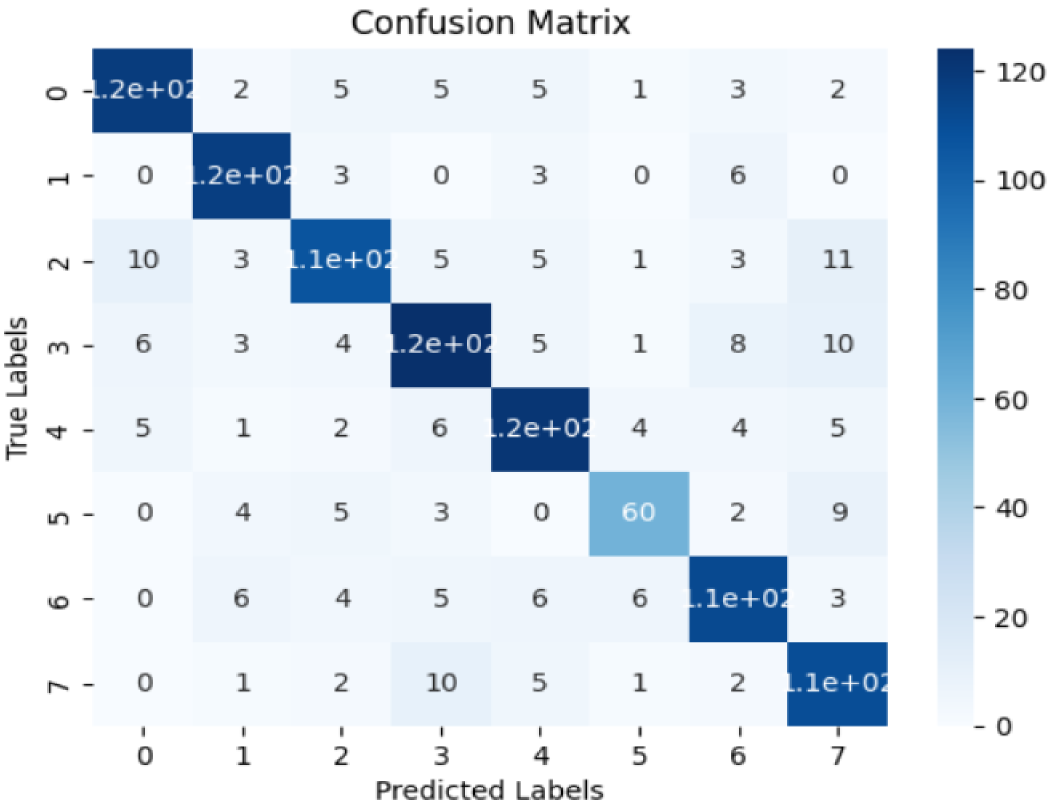
To achieve the goal of emotion detection, the project utilized machine learning techniques and a dataset of speech recordings labelled with corresponding emotions. The dataset was collected from various sources and encompassed a wide range of emotional states, including happiness, calm, surprise, sadness, anger, disgust, fear, and neutral.

The project successfully developed an emotion detection system that could analyse speech inputs and predict the corresponding emotions. However, the overall accuracy and reliability of the system depended on various factors, including the quality of the training data, the choice of machine learning models, and the complexity of the emotions being detected.

The system's potential applications included emotion-aware virtual assistants, emotion-based marketing analytics, mental health monitoring tools, and sentiment analysis in customer service or market research. The project laid the foundation for further research and development in the field of emotion detection through speech analysis, opening avenues for enhanced human-computer interaction and emotional understanding.



	Predicted Labels	Actual Labels
0	Fear	Fear
1	Surprise	Neutral
2	Sad	Sad
3	Neutral	Neutral
4	Fear	Fear
5	Angry	Angry
6	Surprise	Fear
7	Fear	Fear
8	Disgust	Disgust
9	Calm	Calm



	0	1	2	3	4	5	6	7
0	1.2e+02	2	5	5	5	1	3	2
1	0	1.2e+02	3	0	3	0	6	0
2	10	3	1.1e+02	5	5	1	3	11
3	6	3	4	1.2e+02	5	1	8	10
4	5	1	2	6	1.2e+02	4	4	5
5	0	4	5	3	0	60	2	9
6	0	6	4	5	6	6	1.1e+02	3
7	0	1	2	10	5	1	2	1.1e+02

```
Overall Accuracy: 80.46%
```

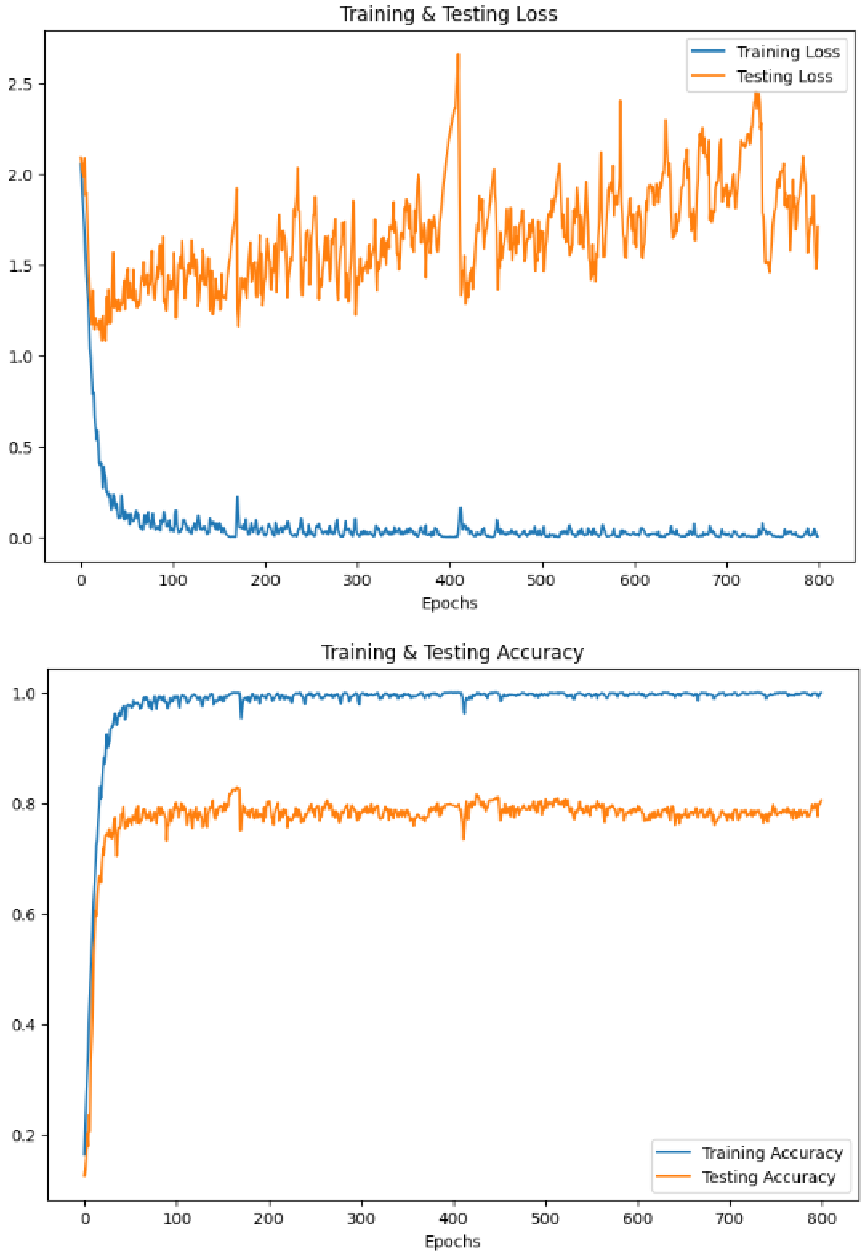


Fig 2 – Output of model

Reference	Algorithm	Accuracy (%)	
[1]	CNN & LSTM	50.22%	
[4]	DCNN	86.3%	
[5]	CNN	60%	
[6]	SVM	81.13%	
[7]	R-CNN	71.5%	

VI. CONCLUSION

Since standard feedforward neural networks cannot handle speech data well (due to lacking a way to feed information from a later layer back to an earlier layer), thus, CNNs have been implemented to take into account the temporal dependencies of speech data. Furthermore, CNNs cannot handle the long-term dependencies due to vanishing/exploding gradient problems very well. Therefore, LSTMs and Bi-LSTM were introduced to overcome the shortcomings of RNNs. This paper evaluated CNN and LSTM.

REFERENCES

[1] S. Zhang, X. Zhao, and Q. Tian, “Spontaneous Speech Emotion Recognition Using Multiscale Deep Convolutional LSTM,” in IEEE Transactions on Affective Computing, vol. 13, no. 2, pp. 680-688, April-June 2022 doi: 10.1109/TAFFC.2019.2947464.

[2] R. A. Khalil, E. Jones, M. I. Babar, T. Jan, M. H. Zafar, and T. Alhussain, "Speech Emotion Recognition Using Deep Learning Techniques: A Review," in IEEE Access, vol. 7, pp. 117327-117345, 2019, doi: 10.1109/ACCESS.2019.2936124.

[3] Selvaraj, Mahalakshmi & Bhuvana, R. & Karthik, S Padmaja. (2016). Human speech emotion recognition. 8. 311-323.

[4] S. Zhang, S. Zhang, T. Huang and W. Gao, "Speech Emotion Recognition Using Deep Convolutional Neural Network and Discriminant Temporal Pyramid Matching," in IEEE Transactions on Multimedia, vol. 20, no. 6, pp. 1576-1590, June 2018, doi:10.1109/TMM.2017.2766843.

[5] K. H. Lee and D. H. Kim, "Design of a Convolutional Neural Network for Speech Emotion Recognition," 2020 International Conference on Information and Communication Technology Convergence (ICTC), 2020, pp. 1332-1335, doi:10.1109/ICTC49870.2020.9289227.

[6] S. Lalitha, A. Madhavan, B. Bhushan, and S. Saketh, “Speech emotion recognition,” 2014 International Conference on Advances in Electronics Computers and Communications, 2014, pp. 1-4, doi:10.1109/ICAECC.2014.7002390.

[7] T. -W. Sun, "End-to-End Speech Emotion Recognition with Gender Information," in IEEE Access, vol. 8, pp. 152423-152438,2020, doi:10.1109/ACCESS.2020.3017462.