Emotion Identification in Speech Using Convolutional Neural Network and Long Short-Term Memory in Deep Learning

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

Guide, Department Of Information Technology[1]
Students, Department Of Information Technology[2,3,4,5]
Zeal College Of Engineering And Research, Pune, Maharashtra, India

Abstract— **The role of emotions in human mental health is extremely important. Since spontaneous emotions in real scenarios are more challenging to detect than other emotions, emotion recognition in real scenarios has attracted a lot of attention in affective computing recently. Motivated by the aid of using the numerous outcomes of various lengths of audio spectrograms on emotion identification, this paper proposes a Long Short-Term Memory (LSTM) model for speech emotion recognition. Initially, a deep convolutional neural network (CNN) version is used to research deep segment-stage capabilities at the foundation of the created image-like 3 channels of spectrograms. Then, a deep LSTM version is followed on the idea of the found-out segment-stage CNN capabilities to seize the temporal dependency amongst all divided segments in an utterance for utterance-stage emotion recognition. Finally, special emotion recognition results, received with the aid of using combining CNN with LSTM at more than one length of segment-stage spectrograms, are included with the aid of using the use of a score-stage fusion strategy.**

***Index Terms*— CNN, LSTM, Emotion and AlexNet.**

I. INTRODUCTION

Emotion recognition in-real scenarios like within the wild have attracted in-depth attention in emotion computing as long as existing spontaneous emotions in-real scenarios square measure more complex and difficult to spot than different emotions.

Driven by the varied effects of different lengths of audio spectrograms in sensation identification, this article proposes a multi-scale Deep Convolution Long Short-Term Memory (LSTM) frame for the spontaneous recognition of speech sensation. Initially, a deep convolutional neural network (CNN) model is employed to determine out deep segment-level options based on the created image-like, three channels of spectrograms.

Then, a deep LSTM model acquires on the premise of the learnt segment-level CNN options to capture the temporal dependency among all divided segments in associate auditory communication for utterance-level feeling recognition. Finally, various sentiment recognition results, obtained by combining CNN with LSTM at multiple spectrogram lengths at the segment level, integrated square measure using a fusion strategy at the score level.

II. PROBLEM STATEMENT

Due to advancements in technology, more and more mental stress is increasing, which is mainly affected by emotions. Therefore, to obtain an accurate emotional state of a person, we will use a machine as it cannot be affected by external factors. Speech being the most used medium of communication, we are detecting emotions through speech.

III. PROJECT OVERVIEW

As human beings' speech is the most natural thing, thanks to specific thoughts. As emotions play an important role in communication within the world, the detection and analysis of emotions are important in today's digital world for remote 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, counselling, 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.

IV. LITERATURE SURVEY:

Since existing 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 a lot of 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 learnt 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 greatly influenced by emotions. It's a way to share one's perspective or state of mind with others. The extraction of the speaker's emotional state from their speech signal is what is meant to be meant by the term "Speech Emotion Recognition" (SER). Any intelligent system with limited computational resources can be trained to identify or synthesize a few universal emotions, such as neutral, anger, happiness, and sadness. Because they both contain emotional information, spectral and prosodic features are used in this work for speech emotion recognition. One of the spectral characteristics is the Mel-Frequency Cepstral Coefficients (MFCC).

Prosodic features such as fundamental frequency, loudness, pitch, speech intensity, and glottal parameters are used to model various emotions. To compute the relationship between speech patterns and emotions, the potential features of each utterance were extracted. The selected features can be used to identify pitch, which can then be used to classify gender. This work uses a Support Vector Machine (SVM) to classify gender. Based on the selected features, the emotions were recognized using the Radial Basis Function and the Back Propagation Network. It has been demonstrated that the radial basis function produces more accurate results for emotion recognition than the backpropagation network does.[3]

The affective gap between subjective emotions and low-level features makes speech emotion recognition difficult. Deep Convolutional Neural Networks (DCNN) have demonstrated remarkable success in bridging the semantic gap in visual tasks such as image classification and object detection by integrating multi-level feature learning and model training. This study investigates the use of a DCNN to close the emotional gap between speech signals. To accomplish this, we first extract three logs of Mel spectrogram channels—static, delta, and delta-delta—that are comparable to the RGB image representation and serve as the DCNN input. After that, the large ImageNet dataset-pre-trained AlexNet DCNN model is used to learn high-level feature representations for each segment divided from an utterance. A Discriminant Temporal Pyramid Matching (DTPM) method is used to combine the learnt segment-level features. DTPM creates a global utterance-level feature representation by combining the temporal pyramid matching with optimal Lp-norm pooling. Linear Support Vector Machines (SVM) are then used to classify emotions. The fact that the DCNN model has been pre-trained for image applications and performs reasonably well in affective speech feature extraction is yet another intriguing finding. 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 rely 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 speech raw data without using any 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]

V. RESEARCH SCOPE:

A system known as speech emotion recognition is used to classify a variety of audio speech files into a variety of different emotions, such as happy, sad, angry, neutral, and so on. In our project, we are trying to detect 4–6 emotions using CNN and LSTM.

Therefore, the research we have done on how to make these models and finding the right datasets suitable for the model we are making.

VI. METHODOLOGY:

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:

LSTM stands for Long Short-Term Memory, employed in the sector of Deep Learning. it's a spread of recurrent neural networks (RNNs) that are capable of learning semipermanent dependencies, particularly in sequence prediction problems. LSTM has feedback connections, i.e., it is capable of processing the whole sequence of knowledge, except for single data points equivalent to images. This finds application in speech recognition, machine translation, and so forth LSTM could be a special reasonable RNN, that shows outstanding performance in an oversized form of problems.

VII. SYSTEM ARCHITECTURE:

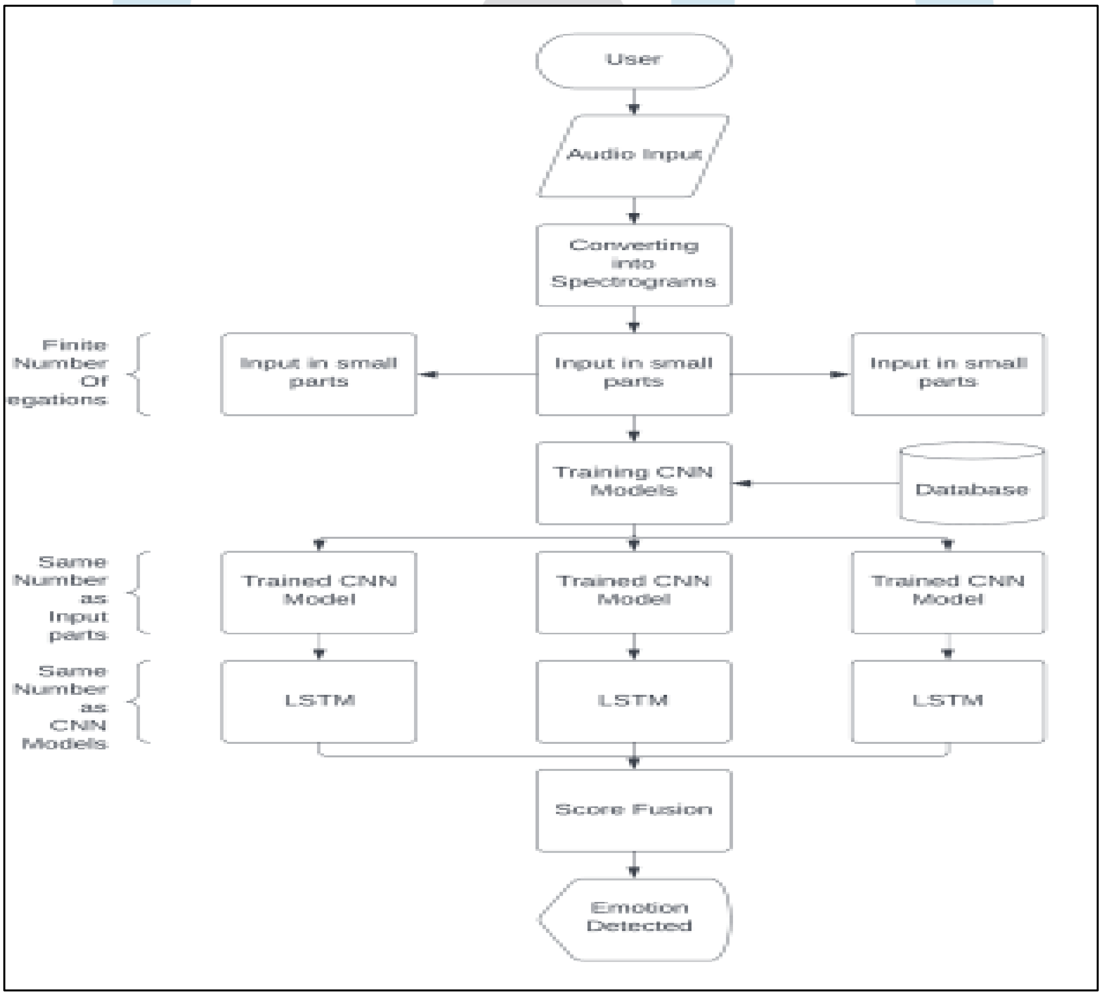

Fig 2: System Architecture

VIII. UML DIAGRAM:

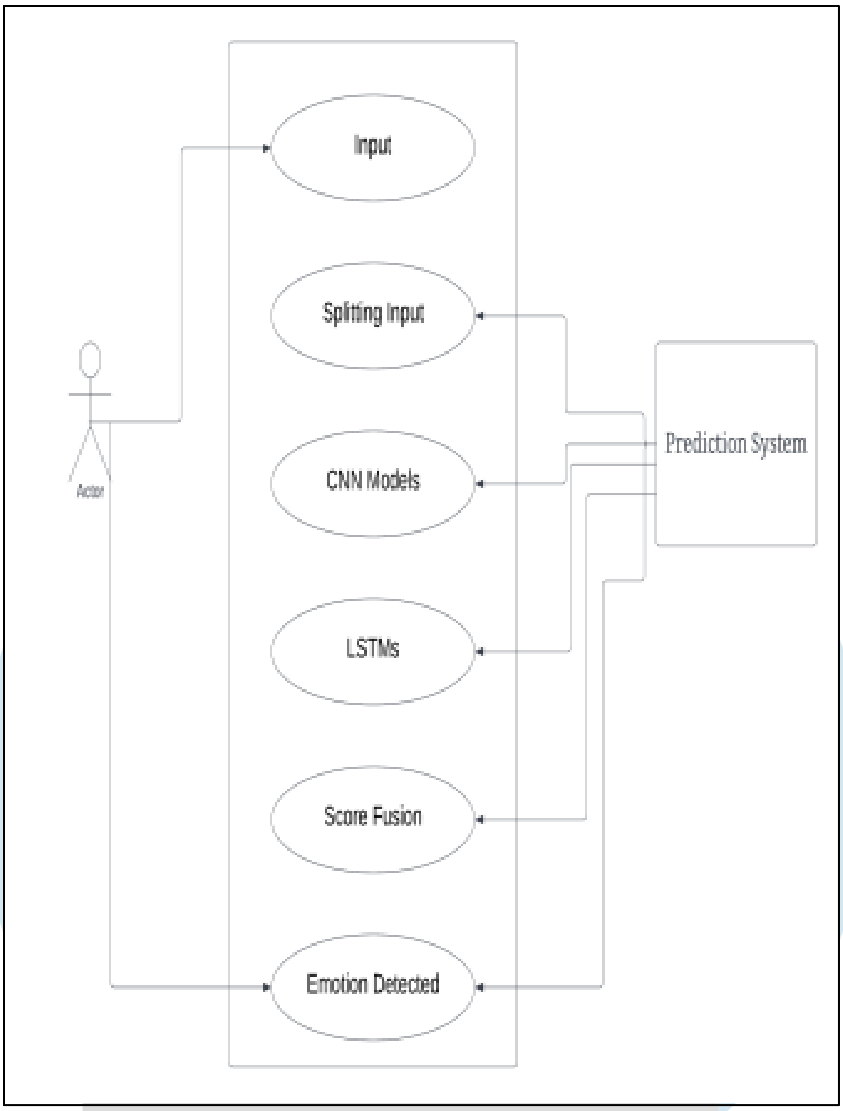

Fig1: UML Diagram

IX. FUTURE SCOPE:

Speech and natural language processing are at the forefront of any human-machine interaction environment today. This highlights the tremendous progress made in machine learning, statistical data mining, and pattern recognition approaches that help makes voice interfaces more versatile and ubiquitous. The growing demand for voice interfaces also warns of potential obstacles to the successful implementation of acoustically robust natural interfaces. Finally, technological advances and research efforts require powerful real-time speech recognition that will completely change the way people interact with their computing devices.

X. 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.