Speech Emotion Detection Using LSTM

1Khushi D M, 2Kopparapu Nihith, 3Umashree

BMS College of Engineering
Bengaluru, India

Abstract - Conceptual Finding the specific feeling of a human in light of speech is exceptionally difficult. Discourse feeling acknowledgment is an undertaking that includes recognizing feelings communicated in discourse. Long momentary memory (LSTM) networks are a sort of intermittent brain network that can be utilized for this errand. LSTMs are appropriate for discourse feeling acknowledgment since they can catch long haul conditions in consecutive information like discourse. In a discourse feeling acknowledgment framework utilizing LSTMs, the discourse signal is first handled to remove highlights like Mel-recurrence cepstral coefficients (MFCCs) that catch significant acoustic properties of the discourse. The highlights are then input into the LSTM organization, which has been prepared to anticipate the close to home condition of the speaker in view of the arrangement of elements. The result of the LSTM is a likelihood dissemination over a bunch of feelings, and the feeling with the most elevated likelihood is chosen as the last expectation.

I. INTRODUCTION

Speech emotion recognition is a rapidly growing field that aims to automatically identify emotions from speech. This technology has a wide range of applications, including human-computer interaction, health care, and market research. Traditional methods for speech emotion recognition rely on hand-engineered features and expert knowledge of audio processing, but recent advances in deep learning have enabled the development of automatic feature extraction methods. One such deep learning method is the use of Long Short-Term Memory (LSTM) networks. LSTMs are a type of recurrent neural network that have proven to be highly effective in a variety of sequence-based tasks, including speech emotion recognition.

II. DEEP LEARNING

Deep learning models are constructed using artificial neural networks with multiple layers, hence the term "deep." These networks are designed to automatically learn representations of data by progressively extracting higher-level features from raw input. The network architecture typically consists of an input layer, multiple hidden layers, and an output layer. Each layer is composed of interconnected nodes, called neurons, which perform computations on the input data.

Training a deep learning model involves a two-step process: forward propagation and backpropagation. During forward propagation, the input data is fed through the network, and the activations of each neuron are computed, resulting in an output prediction. Backpropagation is then used to calculate the gradients of the model's parameters (weights and biases) with respect to a loss function. These gradients are used to update the model's parameters, iteratively improving its performance over time.
III. LSTM

LSTM, short for Long Short-Term Memory, is a type of recurrent neural network (RNN) architecture that addresses the vanishing gradient problem, which can occur when training traditional RNNs on long sequences. LSTM was proposed by Sepp Hochreiter and Jürgen Schmidhuber in 1997 and has since become one of the most widely used architectures in deep learning, particularly for tasks involving sequential data. LSTMs are specifically designed to capture and remember long-term dependencies in sequences by introducing memory cells and gating mechanisms. The key idea behind LSTM is the concept of a memory cell, which can store information over extended time intervals, allowing the network to selectively retain or discard information as needed. The architecture of an LSTM consists of multiple memory cells, each having three main components:

[1] Input gate: This gate determines which parts of the input should be stored in the memory cell. It takes as input the current input and the output from the previous time step and outputs a value between 0 and 1 for each element.

[2] Forget gate: This gate controls what information should be discarded from the memory cell. It takes the current input and the output from the previous time step as input and outputs a value between 0 and 1 for each element.

[3] Output gate: This gate determines what information should be output from the memory cell. It takes the current input and the output from the previous time step as input and produces a value between 0 and 1 for each element.

The input, forget, and output gates allow the LSTM to learn which information is relevant to remember, forget, or output at each time step. The memory cell updates its contents based on the input and forget gates and produces the current output based on the output gate.

IV. TECHNIQUES FOR SER

A. Dataset

Dataset plays an important role in any machine learning and deep learning techniques. It is used to train the model. It helps machines to learn. The model will train with the data and act like a living thing.

In this paper, we collect a dataset of 12600 samples of speeches. The dataset is known as ASVP-ESD. It has 13 variations of emotion contained files. Each audio file is a collection of more than 1 emotion. The duration of all audio files are 11 hours. These audio files are taken from movies, websites, crowds, many people etc.

This file contains speeches, non speeches and only audio files. The dataset contains multi language audio files. Those languages are Chinese, English, French, Russian and others.

B. Feature extraction

In any speech dataset, the files will be in the form of a machine in an understandable way. So that the available data will be converted into machine readable format. So that some models are required for preprocessing the date.

Various types of models are available to convert the data, they are MFCC, BYOL-S, BYOL-A, openSMILE, etc. These techniques are very useful for multilingual datasets for conversion.

In this paper we used the technique called MFCC to convert the audio file into machine language as shown in fig 1.1
C. Model selection

Algorithm selection is the crucial step of all of these stages. In this type of data (SER), few algorithms are there to get the optimized results. Those are Recurrent Neural Network (RNN), Long Short Term Method (LSTM), Conventional Neural Network (CNN), Fully conventional Neural Network (FCN), Attention based FCN etc...

Here we are dealing with the Long Short Term Neural Network (LSTM). It is related to the Recurrent Neural Network (RNN). It will get feedback from the previous cell so that accuracy will be high compared to RNN.

E. Performance measure

The Performance measures are used to know the accuracy of the model and efficiency of it. It is used to find the mistakes and prevent them from happening. In this implementation we have dealt with F1 score, Confusion matrix. The accuracy of the model came as 97%. Usage of Dropouts, transfer learning will increase the accuracy of the model.

V. PROPOSED SOLUTION

The proposed solution is composed of LSTM structure as shown in fig 2. This model will be trained with the given input and feature extraction will take place with the help of MFCC classification method. Here user interactions also take place. He can give input to the system and decide the duration of audio. The proposed system will perform actions on user input and display the output to the user.

![Fig 4. System architecture](image)

The below figure shows fig 5 shows the structure of LSTM. The LSTM model is composed of dense layers and drop-out in it. This drop-out will block the neuron to train the model in the neural network.

![Fig 5. LSTM structure](image)
This will increase the accuracy of the model. The activation function called relu and softmax is used in this model. The softmax function will give the output with the number of classes needed in it. Here 13 numbers of outputs or emotions are needed. So the number of classes given is 13.

VI. CONCLUSION

In conclusion, the utilization of LSTM networks in discourse feeling discovery has demonstrated to be a compelling arrangement. The capacity of LSTM to hold memory and interaction successive information makes it appropriate for discourse feeling acknowledgment undertakings. The consequences of studies show that LSTM-based models outflank customary AI techniques concerning precision and heartiness. As we mentioned before in the problem statement, it is built to detect the emotions of humans based on his speech. Our model successfully finds the emotions of each speech.

Firstly, it takes speech from humans and tests it with a pre-trained saved model. The accuracy of the saved model is 97% and loss accuracy is 27%. Here it predicts the given input. Using argmax we find the current emotion of it. Mel frequency Coefficient Constants (MFCC) has been used for feature extraction. It converts the speech data into 40 column data. Which can be easily trained with the model. The model can reduce its performance measure when it is introduced with a new dataset.

VII. FUTURE WORK

In recent years, deep learning has shown promising results for speech emotion detection. Some future directions for this field include:[1] Multi-modal emotion recognition: Incorporating additional modalities such as facial expressions and body gestures can improve the accuracy of emotion detection.[2] Cross-lingual emotion recognition: Developing models that can recognize emotions in speech across different languages and cultures.[3] Transfer learning: Transferring knowledge learned from large-scale datasets to smaller, domain-specific datasets to improve performance in practical applications.[4] Adversarial training: Incorporating adversarial training to make models robust to subtle variations in speech and more resilient to attack.[5] Context-aware emotion recognition: Incorporating contextual information such as the speaker's demographic information, background, and the situation in which the speech was produced to improve the accuracy of emotion detection. Overall, the field of speech emotion recognition using deep learning has great potential for improving our ability to understand and respond to human emotions in speech.

REFERENCES

