International Journal of Research in Engineering and Science (IJRES)
ISSN (Online): 2320-9364, ISSN (Print): 2320-9356
www.ijres.org Volume 10 Issue 7 ǁ July 2022 ǁ PP. 436-440

Speech Emotion Recognition Using Deep LearningTechnique

Mr. Kenaz K Babu[1]&Mr. ValantoAlappat[2]
[1] M.Tech Student, Dept. of CSE, Thejus Engineering College And [2] Assistant Professor, Dept of CSE, Thejus Engineering College

Abstract
Emotion recognition from speech signals is a very important however difficult element of Human-Computer Interaction (HCI). Within the literature of Speech Emotion Recognition (SER), several techniques have been utilized to extract emotions from signals, as well as several well-established speech analysis and classification techniques. Deep Learning techniques are recently planned as an alternate to traditional techniques in SER. This project presents Deep Learning technique and implements a deep learning technique known as LSTM. We create a LSTM model which is a classification model and we use that model to classify the emotions from the audio input. In this project, we analyze and classify various audio files to a corresponding class and visualize the frequency of the sounds through a plot.

Keywords:*Deep Learning, Recurrent Neural Network, Speech Emotion Recognition, Long Short Term Memory, Mel Frequency Cepstral Coefficient, TESS Dataset.*

Date of Submission: 02-07-2022 Date of acceptance: 14-07-2022

I. INTRODUCTION

Emotion recognition from speech is important component for Human-Computer Interaction (HCI). Speech emotion recognition is a technology that extracts emotion features from computer speech signals, compares them, and analyse the feature parameters and obtain emotion changes. Deep Learning has been considered as an emerging Research field in machine learning and has gained more attention in recent years. Deep Learning techniques for SER have several advantages over traditional methods, including their capability to detect the complex structure and features without the need for manual feature extraction and tuning. Recurrent architectures such as Long Short-Term Memory (LSTM) are much effective in speech-based classification

II. EXISTING SYSTEM

The existing work in this area reveals that most of the present work relies on lexical analysis for emotion recognition. Emotion recognition systems based on digitized speech is comprised of three fundamental components: signal preprocessing, feature extraction, and classification.In the first stage of speech-based signal processing, speech enhancement is carried out where the noisy components are removed. The second stage involves two parts, feature extraction, and feature selection. The required features are extracted from the pre-processed speech signal and the selection is made from the extracted features. Such feature extraction and selection are usually based on the analysis of speech signals in the time and frequency domains. During the third stage, various classifiers such as GMM and HMM, etc. are utilized for classification of these features. Lastly, based on feature classification different emotions are recognized

III. OBJECTIVE

The main aim of the project is to build a model to recognize emotion from speech using libraries and the datasets. The classifier will take audio as an input and the emotions in that audio will be presented as output. Deep Learning techniques will be used to recognize emotions and to create a classifier model that classifies seven different emotions.

IV. SCOPE

The importance of speech emotion recognition (SER) is increasing day by day. Major companies are relying on SER to improve their services. SER is used as performance parameter for conversational analysis. It is used to provide services based on the behavior of customer thereby increasing the profitability. SER can play major role in health industry as it specifies the mental state of a person. It can also play major role in international affairs and politics.

V. PROPOSED SYSTEM

The proposed system uses Deep learning for SER.Deep Learning based Speech Emotion Detection (SER) is designed in this project using Toronto Emotional Speech Set (TESS) dataset. This dataset is trained using Long Short Term (LSTM) model. The model takes audio as input and classifies the emotions in it based on the audio.LSTM networks are a type of RNN that uses special units in addition to standard units. LSTM units include a 'memory cell' that can maintain information in memory for long periods of time. This memory cell lets them learn longer-term dependencies. The existing approaches do not have memory cells. So, usage of LSTM will generate maximum accuracy Features are extracted using Mel Frequency Cepstral Coefficient (MFCC) method.Here in figure 1, the deep learning algorithm used is LSTM, which is a type of Recurrent neural networks which has feedback mechanisms as well.

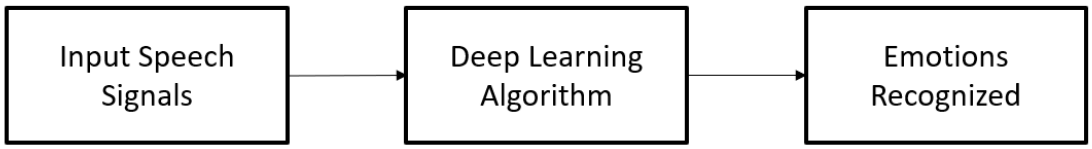

Figure 1: Overall process of proposed system

VI. SYSTEM ARCHITECTURE

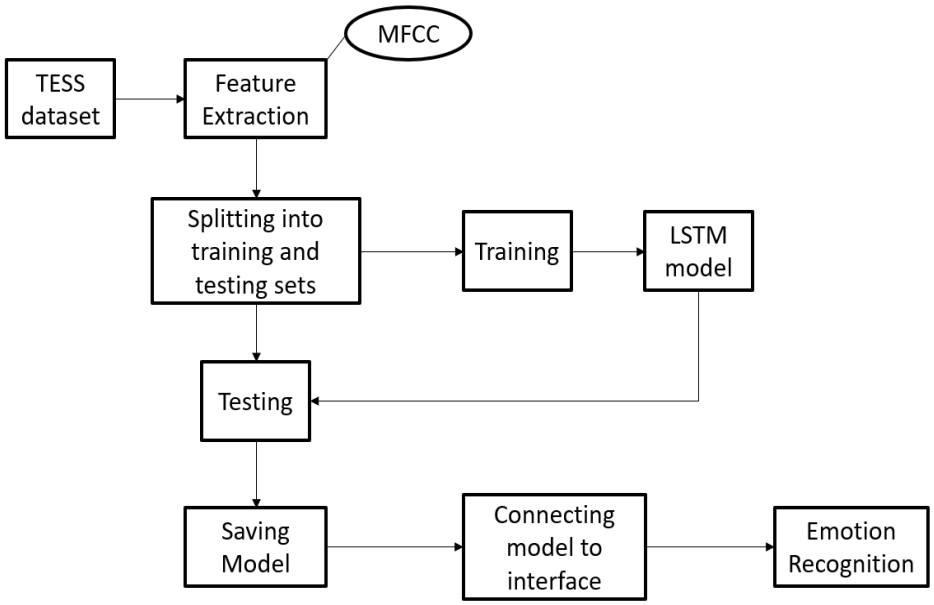

Figure 2: Block Diagram

Figure 2 shows the block diagram of the proposed recognition system. Here, the features from TESS dataset are extracted using MFCC. We divide the dataset with the validation split 0.2 which indicates that 80% dataset is used for training and 20% dataset is used for testing. We train and develop the model by adding number of layers. This model is used for testing the test data. The best accuracy is found out and is saved. Lastly, we connect our model to the interface. The interface will have audio input and the model will give the output.

VII.METHODOLOGY

For the task of speech recognition in this project, we used LSTM model trained on the TESS dataset. First step is to load the dataset. TESS- Toronto emotional speech set data is used for speech emotion recognition.There are a set of 200 target words were spoken in the carrier phrase "Say the word _' by two actresses (aged 26 and 64 years) and recordings were made of the set portraying each of seven emotions (anger, disgust, fear, happiness, pleasant surprise, sadness, and neutral). There are 2800 data points (audio files) in total.

The dataset is organised such that each of the two female actor and their emotions are contain within its own folder. And within that, all 200 target words audio file can be found. The format of the audio file is a WAV format. Output Attributes areAnger, Disgust, Fear, Happiness, Pleasant surprise, Sadness,Neutral. The paths of audio files are loaded for further processing. Filenames were split and appended as labels. To ensure proper processing all characters were converted to lower case. These audio files are preprocessed(duplicates are avoided) and then the exploratory analysis is done. We check for the missing data, and we make sure that all class are in equal distribution. For better understanding, we use wave plot and spectrogram to understand the waveform and frequency distribution respectively. Lower pitched sounds have darker color and high-pitched sounds have brighter color. We do feature extraction using MFCC.

Feature Extraction using MFCC

Speech Recognition is a supervised learning task. In the speech recognition problem input will be the audio signal and we have to predict the text from the audio signal. We can't take the raw audio signal as input to our model because there will be a lot of noise in the audio signal. MFCC is the widely used technique for extracting the features from the audio signal. MFCC stands for Mel frequency cepstral coefficients. There are 4 words in the abbreviation. Mel, frequency, cepstral and coefficients. The idea of MFCC is to convert audio in time domain into frequency domain so that we can understand all the information present in speech signals. Our ear has cochlea which basically has more filters at low frequency and very few filters at higher frequency. This can be mimicked using Mel filters. So, the idea of MFCC is to convert time domain signals into frequency domain signal by mimicking cochlea function using Mel filters.

In MFCC, we will convert our audio signal from analog to digital format with a sampling frequency of 8kHz or 16kHz. we will break the audio signal into different segments with each segment having 25ms width and with the signal at 10ms apart. We will convert the signal from the time domain to the frequency domain by applying the discrete fourier transform (dft). For audio signals, analyzing in the frequency domain is easier than in the time domain. We will use the mel scale to map the actual frequency to the frequency that human beings will perceive and apply log to it. Now we do the inverse dft (idft) transform of the output from the previous step. The periods in the time domain and frequency domain are inverted after the transformations. So, the frequency domain's fundamental frequency with the lowest frequency will have the highest frequency in the time domain. The inverse of the log of the magnitude of the signal is called a cepstrum.The MFCC model takes the first 12 coefficients of the signal after applying the idft operations. Along with the 12 coefficients, it will take the energy of the signal sample as the feature. Along with these 13 features, the MFCC technique will consider the first order derivative and second order derivatives of the features which constitute another 26 features. Derivatives are calculated by taking the difference of these coefficients between the samples of the audio signal and it will help in understanding how the transition is occurring. So overall MFCC technique will generate 39-40 features from each audio signal sample which are used as input for the speech recognition model.

In our project, Audio duration is capped to max 3 seconds for equal duration of file size. It will extract the Mel-frequency cepstral coefficients (*MFCC*) features with the limit of 40 and take the mean as the final feature. We will get a list which is converted to array. It will consist of number of samples and feature. We also use one hot encoder because categorical data has to be converted into numerical data. Our labels are of 7 types and they are categorical. In this method, each category value is converted into a new column and assigned a 1 or 0 (notation for true/false) value to the column.

LSTM model

LSTM is a special kind of recurrent neural network capable of handling long-term dependencies. Long Short Term Memory Network is an advanced RNN, a sequential network, that allows information to persist. It can handle the vanishing gradient problem faced by RNN. A recurrent neural network is also known as RNN is used for persistent memory.The gradients refer to the errors made as the neural network trains. If the gradients start to explode, the neural network will become unstable and unable to learn from training data. Solution to the problem is long short-term memory (LSTM) networks. RNNs built with LSTM units categorize data into short-term and long-term memory cells. Doing so enables RNNs to figure out which data is important and should be remembered and looped back into the network. It also enables RNNs to figure out what data can be forgotten. LSTM has 3 parts. The first part chooses whether the information coming from the previous timestamp is to be remembered or is irrelevant and can be forgotten. In the second part, the cell tries to learn new information from the input to this cell. At last, in the third part, the cell passes the updated information from the current timestamp to the next timestamp.These three parts of an LSTM cell are known as gates. The first part is called Forget gate, the second part is known as the Input gate and the last one is the Output gate**.**Forget gate decides whether we should keep that information. Input gate is used to quantify the importance of new information carried by the input.

For this project we use keras library. We import LSTM, Dense, Droput. We use sequential method and then the model is created. Dense layer is a densely connected neural network layer. Dropout layer adds dropout to input. It is done to prevent overfitting.The Dropout layer randomly sets input units to 0 with a frequency of rate at each step during training time, which helps prevent overfitting. Inputs node set to 0 are scaled up by 1/(1 - rate) such that the sum over all inputs is unchanged. The Dropout layer only applies when training is set to True such that no values are dropped during inference. While compiling we use loss functions, optimizer and accuracy metric The purpose of loss functions is to compute the quantity that a model should seek to minimize during training. Accuracy metric Calculates how often predictions equal labels. Adam optimizers are used to handle sparse gradients on noisy problems. The model is then saved and loaded. We do the validation split 0.2 which means 20% data is used for testing and rest is used for training.

VIII. HARDWARE REQUIREMENTS

1. Processor
    - Laptop - i3 or higher processor
2. Memory
    - Laptop - 4 GB RAM or higher
3. System: PC/Laptop

IX. SOFTWARE REQUIREMENTS

1. Operating system
    - Windows 11
    - Windows 10
    - Windows 7
    - Windows Server 2019
    - Windows Server 2016
    - Linux Distros (Ubuntu, openSUSE, et cetera)
    - Macintosh
2. Programming Language
    - Python
3. Dataset
    - TESS Dataset
4. Computing Environment
    - Jupyter IDE

X. RESULTS AND ANALYSIS

After the model was built, it was saved and loaded. It was used to classify the new audio. We got accuracy of 95%. The implementation results and the respective performance analysis is displayed using screenshots in the following sections.

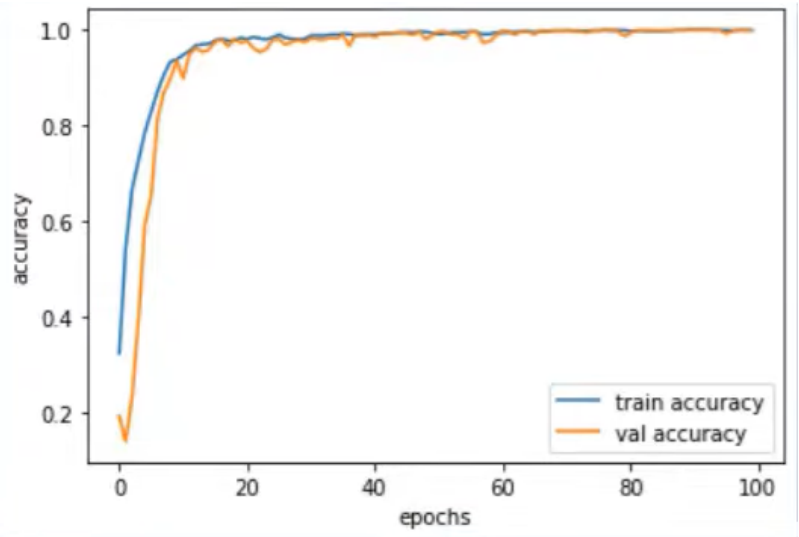

Fig 3: Comparison of train accuracy and validation accuracy

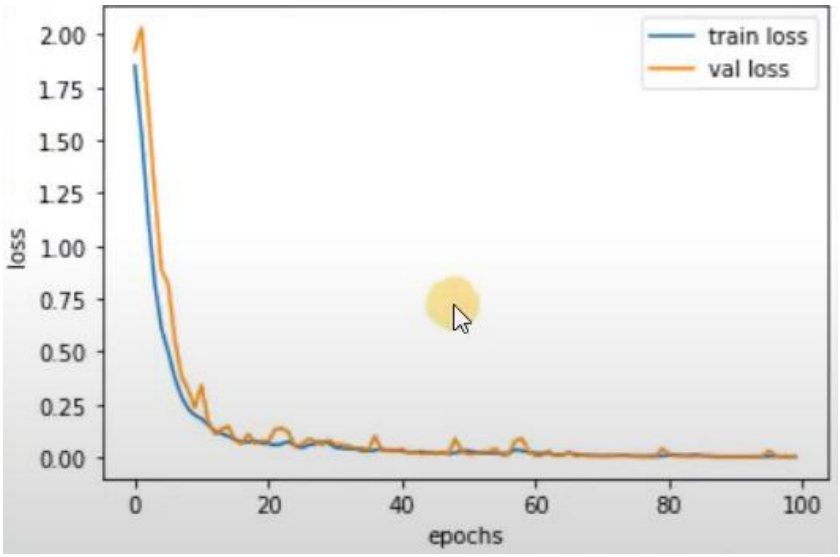

Fig 4: Comparison of train loss and validation loss.

The Performance metrics was analysed both accuracy and the loss was found out. Both the training and testing accuracy shot up to more than 90% after few epochs. The training and validation accuracy was found to be consistent. The losses were there at the initial epochs but after few epochs the loss reduced and nearly reached zero. There were inconsistency at some stage but it stablilizes at the end.

XI. CONCLUSION AND FUTURE ENHANCEMENT

In this project, we proposed a deep learning model that classifies the emotions of the audio based on the input being given. This model is capable of correctly classifying seven human emotions. This project uses a variation of recurrent neural networks called Long Short-Term Memory which helps to classify easily. Fine tuning is done to get maximum accuracy. In future, quality datasets with less noise will help in spot on classification of emotions embedded in speech. Some parameter adjustments can be made to get maximum accuracy. Image based speech emotion recognition can be implemented in the future so that model can predict emotions of people with disabilities like autism and so on. This model can be reused differently depending on the data set and parameters, including speech recognition or other sound related tracks. The system could take into consideration multiple speakers from different geographic locations speaking with different accents.

REFERENCES

[1]. J. Schmidhuber,``Deep learning in neural networks: An overview," *Neural Net.*,vol. 61, pp85_117, Jan.2015
[2]. R. A. Khalil, E. Jones, M. I. Babar, T. Jan, M. H. Zafar and T. Alhussain, "Speech Emotion RecognitionUsing Deep Learning Techniques: A Review," in IEEE Access,vol. 7, pp. 117327-117345, 2019, doi: 10.1109/ACCESS.2019.2936124
[3]. M. S. Hossain and G. Muhammad, ``Emotion recognition using deep learning approachfrom audiovisual emotional big data," *Inf. Fusion*, vol. 49, pp. 6978, Sep. 2019.
[4]. M. Chen, P. Zhou, and G. Fortino, ``Emotion communication system,'' *IEEE Access*, vol.5, pp.326337,2016
[5]. S. Lalitha, A. Madhavan, B. Bhushan and S. Saketh, "Speech emotion recognition," 2014InternationalConference on Advances inElectronics Computers andCommunications,2014, pp. 1-4, doi:10.1109/ICAECC.2014.7002390
[6]. M. S. Likitha, S. R. R. Gupta, K. Hasitha and A. U. Raju, "Speech based human emotion recognition using MFCC," 2017 International Conference on Wireless Communications,Signal Processing and Networking (WiSPNET), 2017, pp. 2257-2260, doi:10.1109/WiSPNET.2017.8300161
[7].] J. Zhao, X. Mao, and L. Chen, ``Speech emotion recognition using deep 1D & 2D CNN LSTM networks,"*Biomed. Signal Process. Control*, vol. 47, pp. 312_323, Jan. 2019
[8]. S. Tripathi and H. Beigi, ``Multi-modal emotion recognition on IEMOCAP dataset using deep learning,"2018, *arXiv:1804.05788*. [Online]. Available: https://arxiv.org/abs/1804.05788
[9]. P.Yenigalla, A.Kumar, S. Tripathi, C. Singh, S. Kar, and J.Vepa, ``Speech emotion recognition using spectrogram & phoneme embedding," in *Proc.Interspeech*, 2018, pp.3688_3692
[10]. E. Lakomkin, M. A. Zamani, C. Weber, S. Magg, and S. Wermter, ``On the robustness of speech emotion recognition for human-robot interaction with deep neural networks," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Oct. 2018, pp. 854_860
[11]. D. Tang, J. Zeng, and M. Li, ``An end-to-end deep learning framework with speech emotion recognition of atypical individuals," in *Proc. Inter-speech*, Sep. 2018, pp. 162_166