IJNRD.ORG ISSN : 2456-4184
INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD) | IJNRD.ORG
An International Open Access, Peer-reviewed, Refereed Journal

SPEECH EMOTION ANALYSIS USING LSTM

1st Yash Khati
Student, Department of CSE, Graphic Era Hill University, Dehradun, Uttarakhand, India khatiyash23@gmail.com

2nd Rakshit Upadhyay Student,
Department of CSE, Graphic Era Hill University, Dehradun, Uttarakhand, India rakshit.upd@gmail.com

3rd Saurabh Kaperwan Student,
Department of CSE, Graphic Era Hill University, Dehradun, Uttarakhand, India saurahkap786@gmail.com

4th Aditya Verma Assistant Professor,
Department of CSE, Graphic Era Hill University, Dehradun, Uttarakhand, India adityaverma@gehu.ac.in

Abstract

Speech emotion recognition is a demanding task in modern day system applications. It is an important research topic that is used to improve public health and contribute towards the ongoing progress of healthcare technology. In current time there are requirements of applications which can work specific task by giving voice commands like Alexa, Google Assistant, Cortana, Siri. But these applications do not recognize human emotion and engage with them. One of the difficult tasks in Speech emotion recognition is to obtain emotion features effectively from user's voice. There has been much research in the field of SER including the use of acoustic and temporal and deep learning models. There has been conducted a lot of research on traditional machine learning algorithms like Support Vector Machine (SVM) [1], K- Nearest Neighbor (KNN) [2], Convolutional Neural Network (CNN) [3], Graph Neural Networks (GNN) [4]. An SER system targets thespeaker's existence by extracting and classifying the prominent features from a preprocessed speech signal. Some primary human emotions are anger, neutral, happiness and sadness, which define the emotional state of human at a particular time which can be classified using trained intelligent system. The improve emotionrecognition accuracy we use features of user voice like pitch, speech intensity and Mel-frequency cepstral coefficients. (MFCC) [5]. Throughout the past ten years, the determination of speech signals emotions was a primary focus but the enhancing the present effectiveness in recognizing needs is imperative, considering the significant dearth of understanding surrounding the fundamental temporal connections inherent in the speech waveform. To fully use the change in emotional content over phase, a new method to voice recognition is now being recommended, integrating structured audio data with Long Short-Term Memory (LSTM) [6] networks. The temporal aspects of the time series were augmented by extracting structural speech features from the waves, now responsible for preserving the intrinsic connections between layers within the actual speech. Many optimized techniques based on LSTM are provided to ascertain emotional concentration across multiple blocks. At the beginning, the approach minimizes computing expenses by altering the traditional forgetting gate. Secondly, instead of relying on the output from the previous iteration of the conventional method, an attention mechanism is used on both the time and feature dimensions within the LSTM's final output. Instead of depending on outcomes from the previous stage, an efficient technique has been used to find the spatial and characteristic aspects in the final output of the LSTM. SER has broad potential in the field of human- computer interaction,healthcare to track The emotional state of patient, providing best user experience through intelligent call centers and bankingsector.

Introduction

Human emotions are difficult to grasp by computer machines. We cannot measure them quantitatively. Speech-based emotion recognition has evolved with the development of machine learning and advanced artificial intelligence algorithms. Human voice features can be extracted in numeric form with the help of various algorithms which can be used to classify human emotions like anger, happiness, sadness, and neutrality. Emotional responses are also an important element of human interactions, it has become an important element of the development of people machine connection-based apps. Facial features, physical sign, and language are various features where science is being used to analyze and interpret reactions. Feeling expressed in audio signal should be continuously identified for seamless interaction between the machine and the user. Human Speech emotions include extraction of features form human speech. To recognize emotions of we must extract paralinguistic information from it.

Since the last decade there have been researchers have suggested various machine learning methods for interpretation of emotion form audio signals. Right now, frustration identification is used to find the satisfaction level of customers. Using this companies personalize their services. Scientists have combined expression identification systems into their application to enhance humancomputer interaction. For making a SER system there are three main steps:

(1). select a good speech database.

(2). find and extract meaningful features.

(3). Create a RNN Model

In SER system power, tone, intensity, amplitude and MFCC are speech characteristics that have speech content. Most researchers use mixed features which are a combination of different characteristics of speech.

Literature Survey

We have explored a lot of research papers with different algorithms like support vector machine (SVM), K-Nearest neighbor (KNN), Convolutional Neural Network (CNN). We focused on deep learning artificial neural network to achieve emotion classification of human voice signal.

[7] CNN with different data augmentation, feature extraction and ensemble learning technique on RAVDESS test dataset to achieve 89.8% accuracy showing effectiveness of these techniques leads to improved generalization capabilityof CNN model.

[8] proposed separate extraction of semantic features using combination of MFCC and Speech2Vec and use of CNN model to transcribe speech spectrogram. The output of speech feature encoder and semantic feature encoder are combined and channeled into LSTM model. The mechanism achieved 94.8% and 94% accuracy on RAVEDESS and EMO-DB dataset respectively.

Another author [9] proposed combination of Bi-directional Long Short Term memory Transformer for extracting speech patterns and 2D CNN for capturing spatial details of audio using mel-spectograms. The output of both models are flattened into 1-D tensors and the resulting vector containing distinct features is sent to dense layer for emotion categorization. The methodology was tested on RAVEDESS and Emo-DB and achieved recognition rate of 95% and 80.19% respectively.

The purposed paper [10] implemented the model using advanced deep learning approach with the help of RNN with a accuracy of almost ninety percentage with condition applied. This model was designed to be used with web applications like automated conversation application with human. This model has some restrictions which the author has described in this purposed paper.

The purposed paper [11] implemented the model of detecting emotions from human voice used extreme machine learning technology. This model leverages the advanced neural network CNN, traditional machine learning algorithm SVM and GMM. Using SVM it achieved approx. 89% accuracy, 90% with CNN model

ad 95% with GMM model.

[12] The purposed paper implemented the model for detecting emotions from speech signals. This mode utilized traditional machine learning algorithms Random Forest model, KNN and MPL All these models achieved approx. accuracy between 70% to 80%. The average accuracy of this model is approx. 84%.

[13] The purposed paper used Hybrid CNN in RAVDESS dataset. It divided emotion into 3 sections for female and male. They proposed the model which have 8 layers of two-dimensional convolutional method. This model has accuracy of more than 98 percentage.

[14] This proposed paper used hybrid CNN-LSTM architecture which have good capability to deal with sequential data. It contains 4 convolutional layers and 3 long short-term memory layers. This model has accuracy around 75%.

[15] The purposed paper solve the problem of lack labelled data for training by data augmentation. It proposes cycle-GAN for data augmentation. These networks are used to produce feature vectors like training set. It uses Wasserstein divergence to overcome vanishing gradient problem. This model hasa accuracy of 83%.

Data Set Collection:

It is important to have diverse and high- quality dataset for data related tasks. It should be huge dataset to avoid chances of overfitting problem. Various type of dataset are available on various web sources. Various publicly available dataset are :

- Ryerson audio-visual database Of Emotional Speech And Song (RAVEDESS)
- Toronto Emotional Speech Set (TESS)
- Crowd Sourced Emotional Multimodal Dataset (CREMA-D) , etc.

RAVEDES dataset is used in this project. It contains 1440 audio files of 24 actors , half men and half female. Most of actors have different recorded emotions happy , sad, neutral , surprise , fearful and disgust. The tabular dataset representation of RAVEDES data :

present for feature extraction, we are using MFCC technique to extract these features.

SNO	Emotion	No Of Files
1	Calm	192
2	Happy	192
3	Sad	192
4	Angry	192
5	Fearful	192
6	Disgust	192
7	Surprised	192
8	Neutral	96

Table 1 : Audio Dataset

Audio Processing:

Before feature extraction from audio we have to preprocess audio to remove noises and silence present the audio files. Audio files contains many silence part which does not provide any information therefore we will remove it from the files. Various methods are available to remove these silence part from the files. One of the ways in deep learning is to use librosa package which is used to read audio file and analyze them. In next step are going to only retain higher frequency and leave all lower frequency because higher frequency contains actual information about speech features.

Feature Extraction :

Speech signals vary with time and its characteristics also change with time. We can visualize the speech signals using amplitude spectrum. Using these short term spectrum we can extract features form the respective speech. Speech signals vary from person to person therefore it is difficult to extract features from the spectrum. There are many techniques

Mel Frequency Cepstral Coefficient (MFCC)

MFCCs are a small collection of unique features mostly used in voice speech processing. It is derived from voice signal to get it's spectral properties like human hears. For feature extraction at first we perform window function on the signal and afterward we apply discrete Fourier transform after computing discrete Fourier transform (DFT) we take the logarithm of magnitudes, set frequencies on Mel Scale, and at last we perform Inverse Discrete Cosine Transform.

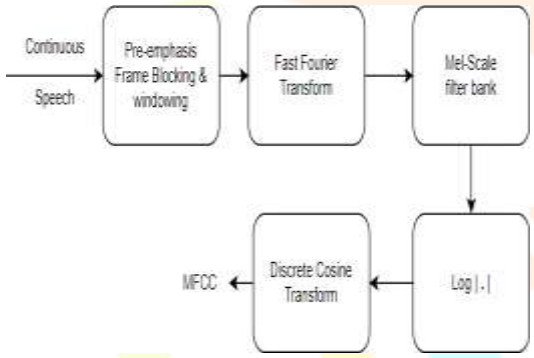

Figure 1: Block Diagram Of MFCC Feature Extraction

Framing :

It divide voice signals into frames each of which last twenty-thirty milliseconds. Since speech signal vary time to time that require framing however it's feature are changed as result spectrum analysis only be done particular length of time.

Windowing :

To contain consistency of voice signal each of the subsequent frame is multiplied by hamming window functions this window function is used to minimized the inconsistency.

P(a) = Q(a) * R(a)

Where R(a) : Function of hamming window

Fast Fourier Transform :

It is used to change frequency domain to time domain. The magnitude frequency of each frame retrieved using fast Fourier transform (FTT). Output of FTT is spectrum which is also known as periodogram.

Algorithm Used:

Long Short-Term Memory (LSTM) model is type of Recurrent Neural Network (RNN) introduced to overcome vanishing gradient problem experienced when training RNN model of large sequence of data. LSTM architecture maintains cell state that helps toremember value of random intervals for long durations. The cell state and hidden state present in LSTM model help model to learn from its previous prediction and error for sequential data. The LSTM cells and three gates that are input, output and forget gate help model to monitor the f vercome vanishing gradient problem.

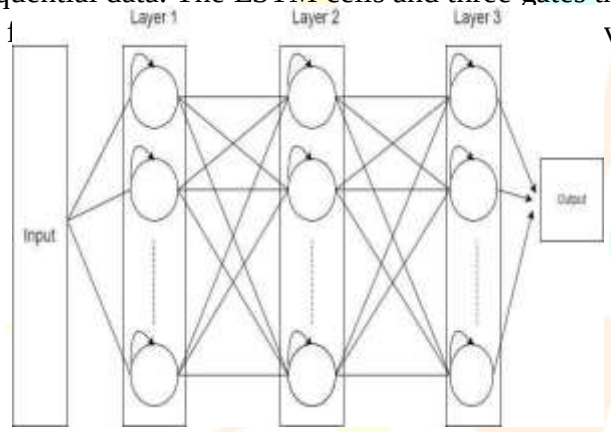

Figure 2: LSTM Architecture

Building Model:

LSTM model is sequential model, and it is built using Keras in python. There are some steps to develop this model.

Defining Network:

In Kera's we define neural networks as a series of layers. Kera's provide high level API to built neutral network. The sequential class serves as framework for these layers. First we built sequential class instance then build a series of layers which are sequentially interconnected in a particular order.

Compiling Model:

After successfully building the model, compile it. It produces an optimized set of matrix transformation which then runs on CPU. Before compiling the model define a set of parameters like optimizer, kernel, anderror function.

Fitting Model:

After building the model it needs to be trained using dataset to adjust the weight of model. Which includes finding the error using error function and adjust weight through optimizer and it is known as back propagation. We perform back propagation a certain number of times which is called number of epoch.

Result and Conclusion

Using LSTM (Long Short-Term Memory) we were successful in implementing the purposed framework Emotion "Analysis Using LSTM" with good precision and accuracy. Implementing MFCC short term spectrum of sound signal with LSTM results in better accuracy. We tested the purposed SER model on thousands of inputs and found excellent results. Using LSTM model, we found better results compared to traditional machine learning algorithms like Support Vector Machine (SVM), K-Nearest Neighbor (KNN) and Convolutional Neural Network (CNN). The purposed SER model will be able to predict more accurate emotions of patients in healthcare, In the business world it can be used in providing better customized customer services according to their emotional state.

References :

[1] Fatima A.Hameed , Dr. Loay E. George "Using Speech Signal for Emotion Recognition UsingHybrid Features with SVM Classifier" 2023 , Waist Journal of Computer and mathematics Science.

[2] Mohammed Ehsab, Eyad I, AbbasSpeech Recognition Algorithm in a Noisy Environment Based on Power Normalized Cepstral Coefficient and Modified Weighted-KNN , 2023 , Engineering and Technology Journal

[3] Mohamed Daouad , El Wardani Dadi , "Amazigh Speech Recognition Using 1D CNN",2023

[4] Georgios Skoulidis, "Speech Emotion Recognition with GNNs", May 2002, Electrical Engineering and Computer Technology DepartmentUniversity Of Patras.

- **[5]** Hongxing Zhang, Hu Li, Wenxin Chen, Hongjun Han, "Feature Extraction of Speech Signal Based on MFCC (Mel cepstrum coefficient)", September 2023, IOP Publishing.

[6] Xun Chen, Chengqi Wang, Yuxin Li, Chao Hu, Qin Wang and Dupeng Cai, "Research on Bidirectional LSTM Recurrent Neural Network in Speech Recognition", June 2023, International Conference on Pattern and Artificial Intelligence.

[7] Avvari Pavithra, Sukanya Ledalla, J. Sirisha Devi "Deep Learning-based Speech Emotion Recognition: An Investigation into a sustainably Emotion-Speech Relationship" 2023 E3S Web of Conferences

[8] Llkhomjon Pulatov, Rashid Oteniyazov, Makhumudov Fazliddin, Young-Im Cho "Enhancing Speech Emotion Recognition Using Dual Feature Extraction Encoders" . Sensors , 2023

[9] Sera Kim , Seok-Pil Lee "A BiLSTM– Transformer and 2D CNN Architecture for Emotion Recognition from Speech" electronics, 2023

[10] Ján Paralic, Jan Micko, Kristína Machová and, Martina Szabóova, "Detection Of Emotion By Text Using Machine Learning" September 2023, Frontiers in Psychology.

[11] Valli Madhavi Koti1, Krishna Murthy, M Suganya, Meduri Sridhar Sarma,Gollakota V S S Sasikumar, Mr Balamurugan N, "Speech Emotion Recognition Using Extreme Mahine Learning" November 2023 EAI Endorsed Transactions.

[12] Roy D Gregori Ayon, Md. Sanaullah Rabbi, Umme Habiba, Maoyejatun Hasana,"Bangla Speech Emotion Detection Using Machine Learning Ensemble Methods", November 2022 ASTES.

[13] Tanvi Puri, Mukesh Soni and Gaurav Dhiman, "Detection of Emotion of Speech for RAVDESS Audio Using Hybrid Convolution Neural Network", February 2022 Hindawi.

[14] Qianhe Ouyang, "Speech emotion detection based on MFCC and CNN-LSTM architecture", 2023 Applied and Computational Engineering.

[15] Arash Hossein Khosravi, Wenwu Wang Shilandari, Hossein Marvi, "Speech emotion recognition using data augmentation method by cycle-generative adversarial networks" , 2022 Springer Nature.