Sentiment Analysis of Human Speech using CNN and DNN

[1]Saumya Roy ,University of Enginneering and Management
Kolkata, India
s4aumyaroy@gmail.com

[2]Sayak Ghoshal University of Enginneering and Management
Kolkata, India
s.ghoshal226@gmail.com

[3*]Rituparna Basak University of Enginneering and Management
Kolkata, India
*rituparna.basak02@gmail.com

Abstract: Communication is the exchange of thoughts, ideas and feelings through emotion. In this paper we have proposed a method where human speech is converted into digital input. The digitized sound is fed into the proposed models and the voice of every person is classified into discrete emotional characteristics by its intensity, pitch, timbre, speech rate and pauses. In the proposed method, authors have applied multi scale area attention in a deep 2D-CNN connected to dense DNN to obtain emotional characteristics with wide range of granularities and therefore the classifier can predict a wide range of emotions on a broad scale classification.

Keywords— Sentiment Analysis, Audio Analysis, Deep Learning, Neural Networks, Emotion Detection, Deep Neural Network(DNN), Convolutional Neural Network(CNN).

I.INTRODUCTION

Speech is considered to be the most valuable and widely used means of communication. Speech Emotion Recognition (SER) has wide application perspectives on psychological assessment, robotics etc. For example, a doctor treating a patient suffering from depression can keep a track of his patient's development and design a recovery plan according to the emotions hidden in the patient's speech. Over the past few years, there has been a consequential development in the field of analyzing the emotions of speech with Deep Learning, but there are still deficiencies in the research of SER, such as insufficient model accuracy, shortage of useful dataset, and lack of computing resources. In SER, emotion may depict distinct energy patterns in spectrograms with varied granularity of areas. However, typical attention models in SER are usually optimized on a fixed scale, which may limit the model's capability to deal with diverse areas and granularities.

In SER, change in energy patterns in spectrograms results in different types of emotions. Typically, an attention neural network[1] classifier of speech emotion recognition is usually optimized on a fixed attention granularity. A drawback of an attention neural network classifier of SER is that it is usually optimized on a fixed attention granularity. In the proposed method this constraint is removed by applying multi scale area attention in a Deep CNN as well as Dense DNN to obtain emotional characteristics with a wide range of granularities and therefore the classifier can predict a wide range of emotions on a broad scale classification. For example, sentiments such as annoyance, joy have different levels of intensity, so instead of just categorizing the presence or absence of the emotion, level of intensity of the emotion can be identified as well. Hence comparative study using both the models is conducted.

Models that have been used for SER previously suffer a problem of sample scarcity. A novel approach is used to deal with data sparsity. For this reason, augmentation of data with addition of stretching, pitch modification and noise insertion. This adds variation to the dataset and it also improves the chances of getting better accuracy. For example, a training data with all anger emotions expressed in higher pitch will now have the same emotion in lower pitch, hence maximizing the chances of identifying anger, when spoken in a lower timbre. Similarly, the sample set might consist of audio of fast speakers and be essentially biased. Stretching helps in removing this bias. Addition of noise to the data is proven to be a useful[2] tool for classifying real time date, since real time data tend to have background noise. To the effectiveness of the proposed method, extensive experiments are carried out on RAVDESS[8], CREMA-D[9], TESS-D[10] dataset.

.

II.Literature Survey

In the paper [1] ,authors have studied a new techniques of utterance-based emotion recognition. The comparison between the efficiency of Support Vector Machines (SVM) and Binary Support Vector Machines (BSVM)[14] techniques are clearly depicted by the authors. The different frame based features like,Acoustic features including energy, MFCC[10][11], Perceptual Linear Predictive[12][13] (PLP), Filter Bank (FBANK)[15], etc. are taken into considerations.

In the paper [2] an improved multimodal approach for sentiment analysis is proposed. The basic goal for this paper is binary classification of sentiments that is either positive or negative. For a better user experience, from the speaker speech emotion, age or gender was recognized. In this paper, the technique of Two-dimensional Convolutional Neural Networks [16] and Deep Neural Networks is used for encoding each segments into a vector of fixed length by integrating the activations of the last hidden layer over time.

In the paper [17] proposes a real-time Speech emotion recognition system based on End-to-end (E2E) learning.From a one second frame of raw speech spectrograms the technique of deep neural network is used to study the emotions. A deep hierarchical framework, pragmatic optimization and data augmentation helps in achieving the desired results. Promising results are reported.

In the paper [3] a well organised procedure has been provided by the author for implementing SER political debates; The emphasis is laid on manufacturing the outcome and then to prepare visualisation of the said results. Two alternative approaches have been considered, a classification-oriented approach and a lexicon-oriented approach. In the former universal and domain oriented sentiment lexicons are used. Two general techniques for implementing domain oriented lexicons-based approach has also been considered. These are (a) direct generation and (b) adaptation. Direct generation focuses on producing exclusive lexicons depending upon the data labels. Adaptation considers a common and inclusive lexicon based approach and adjusting it as per necessity to develop it into a non-generic and exclusive symbol of a particular domain.

The results obtained from the above discussed approaches were considered and compared with the “classification-based” approach. By observing and analysing the attitude of the political speakers in the debates, the sentiment mining approaches were compared. Collective labelled speech data was considered, which were of political significance which was extracted from debating transcripts. The outcome of the comparison helped them realise that using sentiment mining the speakers attitude can be determined conclusively.The proposed Debate Graph Extraction (DGE) framework, in its functioning, effectively extracts the debate graphs from political debate transcripts. They proposed to graphically represent debates with speakers as nodes. In this framework, the speakers are represented as nodes, with nodes having specific labels and links between nodes. These links depend upon the exchange of speeches. The labels on the nodes depended upon the sentiment of the speakers. The attitude of the speaker was then used to classify a link as supporting or non-supporting. If the outcome of both speakers was same, i.e. both positive or both negative then the link was labelled as supporting or else it was labelled opposing. Visualisation of results were carried out via graphs that represent the essence of the debate, in an abstract manner. Lastly they discuss about how debate graphs can be structurally analyzed using the techniques based on network mathematics and community detection techniques.

In the paper [4] ,they proposed automatic sentiment detection system for natural audio streams. Part of speech tagging and maximum entropy modelling (ME) has been used as the suggested technique to develop a sentiment detection model, that was text-based in nature. The number of model boundaries in ME was reduced drastically by an attuning technique while conserving the classification capability. Using decoded ASR (automatic speech recognition) transcripts and the ME sentiment model, sentiments of YouTube videos were able to be determined. As evaluation, they have gathered motivating classification accuracy. According to the results analysis showed that performance on sentiment analysis on spontaneous speech data is possible in spite of word error rates.

In the paper [6] aims to underline different techniques to detect vocal expressions of different emotional states. Moreover, the features extracted from machine learning methods and speech datasets were analysed with an emphasis on classifiers. Additionally, this paper shows the outline areas where emotion recognition can have an effective application like cognitive sciences , psychology, marketing and healthcare.

III.Problem Statement

The human speech is the most natural way of expressing ourselves. We know emotions play an important role in communication analysis, and the detection of the same is significantly important in today’s digital world of remote communication. In text based classification certain emotions like sarcasm, dual meaning sentences cannot be identified. Tonal Qualities of the voice is required to classify the emotions more accurately. An SER system can be defined as a collection of methodologies that classifies speech signals to detect embedded emotions. The human speech contains many features different to each individuals. If we consider all those features while training the model, then the model will be biased to the training set which is not desired. So we have considered only the properties common to human voices like loudness, timbre, and quality. Our attempt lies in trying to detect underlying emotions embedded in speech through analysis of the acoustic features of the audio recording.

IV. Dataset

Three instances of audio datasets was used during our analysis, which contains the vocal emotional expressions in sentences spoken in a range of (joy, grief, rage, agitation, annoyance, and calm). Total 1440 and 7,440 clips of 115 actors were collected which had a diverse ethnic background, it was merged with 8882 files. We have worked only with the audio recordings of the audio-visual data. The sentences are spoken by trained Actors belonging to a variety of races and ethnicities (Latino Americans, African American, Asian, Caucasian). The sentences are classified using one of six different emotions (Joy, Grief, Rage, Agitation, Annoyance, and Calm) and four different emotion levels (Low, Medium, High, and Unspecified). The audio file format is WAV. We have preprocessed the dataset to clear noise and to introduce stretching.

V. Proposed Solution

Three classes of features can be mainly identified in a speech. These can be classified as lexical features, the visual features, and the acoustic features. For example: the various expressions of the speaker, the terminology used, and properties like vocal quality, pitch, anxiety, noise, energy, etc.

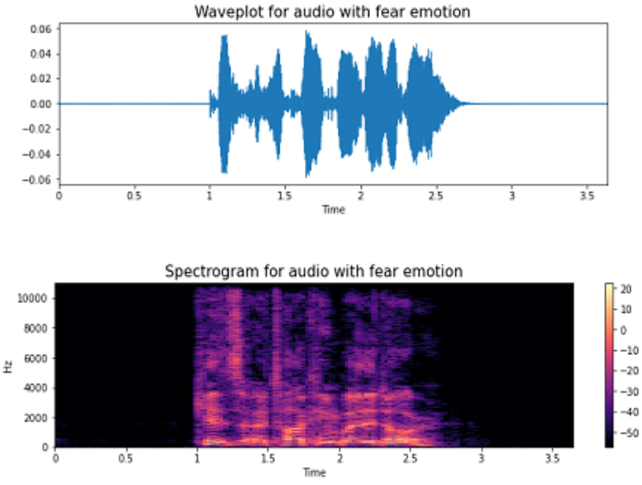

Fig 5.1. The spectrogram for audio with fear emotions.

Analysis of lingual features would require a script of the speech. However, it will require a processing and extraction of text from speech in order to analyze sentiments from real-time audio. Analyzing visual features would require the access to the video of the conversations, and it is not in the scope of this research. Therefore, analysis of the auditory features is done in this work, since analysis of the acoustic features is possible in real-time. Audio from real-time conversations is extracted and analysed in order to accomplice the task

Furthermore, the representation of emotions can be done in two ways:

- Discrete Classification: Emotions were classified into distinct labels like rage, calmness, neutral, cheerfulness and joy etc.
- Dimensional Representation: Representations of emotions with dimensional categories such as Activation Energy (on a low to high scale), Valence (on a negative to positive scale), or and Dominance (on an active to passive scale)

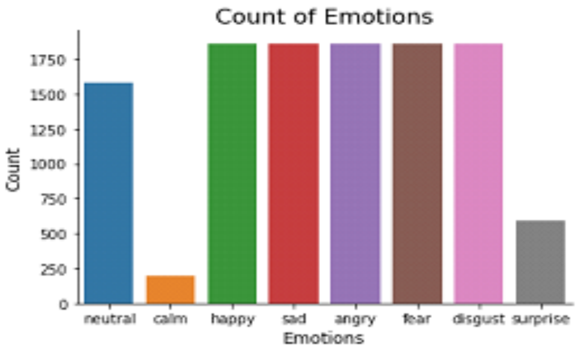

Fig 5.2. The various categories of sentiments our model predict

The two mentioned approaches have their distinct advantages and disadvantages. The dimensional representation approach is an elaborative process but there is a lack of annotated audio data in the dimensional format. The discrete classification is more straightforward and less resource hungry to implement. In discrete classification approach, emotions are classified on a specified scale for the analysis. Emotions are classified using the trained model and predicted as discrete outputs. This approach is easier to implement and understand and as such has greater outreach.

In the proposed method , discrete classification approach is used for analysis. We classify the emotions on a specified scale. The emotions are classified using the trained model and emotions are predicted as discrete outputs. This approach has a greater outreach.

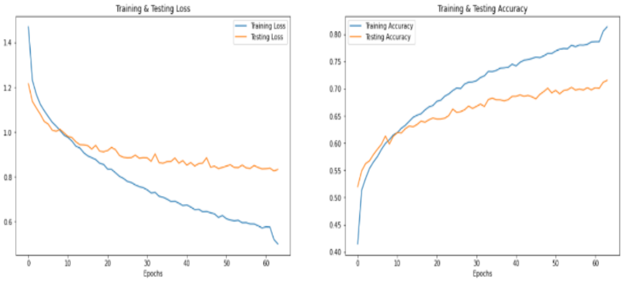

Fig 5.3. (a) Training and Testing loss **(b)** Training and Testing
Accuracy of 2D-CNN model 2D- CNN model

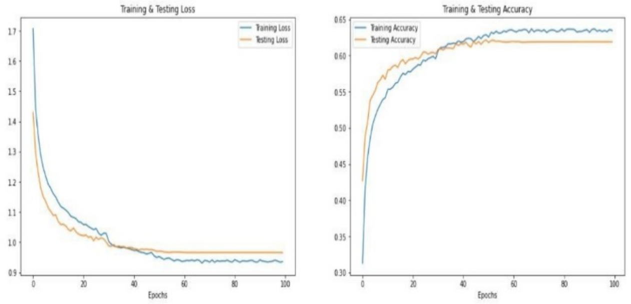

Fig 5.4. (a) Training and Testing loss **(b)** Training and Testing
Accuracy of Dense DNN model Dense DNN model

Figure 5.3 shown above plots the output of the training and testing for 2D-CNN model. In the X-axis, the number of epochs is plotted.

In Fig. 5.3(a) training and testing loss of 2D-CNN model is shown and Fig. 5.3(b) training and testing accuracy of 2D-CNN model is shown.

On analysis of the graphs, each epoch in both the graphs tend to merge at a point which is desired. After the point of saturation, there is no significant change in the difference between training and testing loss of the model and also training and testing accuracy of the model which shows the strong integrity of the proposed model.

Figure 5.4 shown above plots the output of the training and testing for the Dense DNN model. In the X-axis, number of epochs is plotted.

In Fig. 5.4(a) training and testing loss of Dense DNN model is shown and Fig. 5.4(b) training and testing accuracy of Dense DNN model is shown. On analysis of the graphs, with each epoch in both the graphs the lines tend to merge at a point. After merging, the lines diverge, which indicates significant variance between the training and the testing loss. With higher number of epochs, the loss decreases since the model refuses to achieve saturation. In case of the accuracy plot , with higher number of epochs, the accuracy increases since the model refuses to achieve saturation.

After careful observation and improvement of our model, we are able to achieve an accuracy of 61.89%. Thus, our model is able to provide some noteworthy results that can have myriad of applications.

VI.CONCLUSION

In the proposed method, an accuracy of 61.89% is achieved using both DNN and CNN. Thus, our model is able to provide some noteworthy results that can have wide range of applications. Efficient utilization of the audio signals and their tone, pitch and granularity can also help in detection of lies, mimicry as well as mental state of a person. A text-based approach will not provide such pronounced outcomes as they are bound only to linear degree of variation. Furthermore, for exploring broader spectrum analysis and interpretation, such as analysis of interviews, interrogations etc., Multimodal Sentiment Analysis should be taken into considerations.

For future work, we intend to expand and add more features into the proposed framework. Bidirectional LSTM is also a convenient proposition for training over audio datasets MFCC. Addition of embedding framework, as attention frameworks[7] seem to work well for many voice recognition tasks, or residual layers when there is an absence of handful labels, and there is a high possibility of overfitting. Collecting more data in the feature, as TESS and RAVDESS[8] only provide limited samples of user information is also a direction to be explored and analyzed. Due to limited data augmentation of data and variation such as using stretch and adding noise has been done where most of those features would not be impacted. The application of sentiment analysis techniques can be used to predict the demeanor of individuals. Increasing the spectrum in sentiment classes may provide valuable information, which is not captured efficiently earlier. Our efforts should also highlight the tendency of our model to isolate hateful speeches and sexist remarks.

VII.REFERENCES

[1] Wenpeng Yin, Hinrich Schütze, Bing Xiang, Bowen Zhou; ABCNN: Attention-Based Convolutional Neural Network for Modeling Sentence Pairs. Transactions of the Association for Computational Linguistics 2016; 4 259–272. doi: https://doi.org/10.1162/tacl_a_00097t

[2] Roland Goecke; Gerasimos Potamianos; Chalapathy Neti, "Noisy audio feature enhancement using audio-visual speech data" DOI: 10.1109/ICASSP.2002.5745030

[3] Nattapong Kurpukdee , SawitKasuriya , VatayaChunwijitra ,Chai Wutiwiwatchai and PoonlapLamsrichan ,"A Study of Support Vector Machines for Emotional Speech Recognition", 978-1- 5090-4809-0/17/$31.00 ©2017 IEEE

[4] HarikaAbburi, "Audio and Text based Multimodal Sentiment Analysis using Features Extracted from Selective Regions and Deep Neural Networks", International Institute of Information Technology Hyderabad - 500 032, INDIA June 2017

[5] Zaher Ibrahim Saleh Salah, "Machine Learning and Sentiment Analysis Approaches for the Analysis of Parliamentary Debates", May 2014

[6] Suraj Tripathi, Abhay Kumar, Abhiram Ramesh, Chirag, Singh, Promod Yenigalla, "Deep Learning based Emotion Recignition System Using Peech Features and Transcription"

[7] LakshmishKaushik, AbhijeetSangwan, John H. L. Hansen," Sentiment Extraction From Natural Audio Streams", 978-1- 4799-0356-6/13/$31.00 ©2013 IEEE

[8] S. Lugović, I. Dunđer and M. Horvat,"Techniques and Applications of Emotion Recognition in Speech", MIPRO 2016, May 30 - June 3, 2016, Opatija, Croat

[9] Omer Tai, Yang Liu, Jimmy Huang, Xiaohui Yu, Bushra Aljbawi., "Neural Attention Frameworks for Explainable Recommendation".

[10] Mingke Xu 1 , Fan Zhang2 , And Wei Zhang "Head Fusion: Improving the Accuracy and Robustness of Speech Emotion Recognition on the IEMOCAP and RAVDESS Dataset" DOI 10.1109/ACCESS.2021.3067460

[11] Houwei Cao, David G. Cooper, Michael K. Keutmann, Ruben C. Gur, Ani Nenkova, Ragini Verma "CREMA-D: Crowd-Sourced Emotional Multimodal Actors Dataset"

[12] Muhammad Zafar Iqbal "MFCC and Machine Learning Based Speech Emotion Recognition Over TESS and IEMOCAP Datasets".

[13] Fang Zheng, Guoliang Zhang & Zhanjiang Song "Comparison of different implementations of MFCC".

[14] Hynek Hermansky "Perceptual linear predictive (PLP) analysis of speech".

[15] Hynek Hermansky, Louis Anthony Cox, Jr "Perceptual Lnear Predictive (Plp) Analysis-Rsynthesis Technique".

[16] Sungmoon Cheong, Sang Hoon Oh, Soo-Young Lee "Support Vector Machines with Binary Tree Architecture for Multi-Class Classification".

[17] Martin Vetterli, Cormac Herley "Wavelets and Filter Banks: Theory and Design".

[18] Jing Chang, Jin Sha "An efficient implementation of 2D convolution in CNN".

[19] H.M. Fayek; M. Lech; L.Cavedon; "Towards real-time Speech Emotion Recognition using deep neural networks" DOI: 10.1109/ICSPCS.2015.7391796.

[20] Ziquan Luo; Hua Xu; Feiyang Chen; "Audio sentiment analysis by heterogeneous signal features learned from utterance-based parallel neural network" DOI: 10.1109/ICSPCS.2015.7391796.