International Journal of Engineering Sciences & Emerging Technologies, May 2024.
ISSN: 22316604 Volume 11, Issue 3, pp: 628-632 ©IJESET

SPEECH EMOTION RECOGNITION SYSTEM: A REVIEW

Priyanka Goel, Shubham Yadav, Vasundhra Gupta,

Zareen Aqiq, Shrey Ruhela

Assistant Professor Department of CS & E
Moradabad Institute of Technology,Moradabad, UP, India
priyanka070goel@gmail.com

CS&E, MIT, Moradabad
Moradabad Institute of Technology,Moradabad, UP, India
shubhamyadav653@gmail.com
guptavasundhra2903@gmail.com
shreyruhela1311@gmail.com

ABSTRACT

Understanding the human emotional state has been extensively used in various applications. Emotion recognition has been used extensively in human computer interaction systems to derive important features from speech. Emotional state of a human being is mainly influenced by physical characteristics such as muscle tension, skin elasticity, and blood pressure. The emotions of a person are unique in nature but their understanding, interpretation, and reflections can be distinct.

KEYWORDS

Emotion recognition, feature extraction, pre-processing, MFCC, classifier

1. INTRODUCTION

Human emotions are considered as mental and physiological state. Detecting human emotions can play an important role in intelligent human computer interaction system. Emotions can also be recognized through the waya person speaks.

But at the same time, this is very challenging task as it is difficult to predict the emotional state of a person using the audio obtained through his/her speech. Emotion recognition has various utilities in fields such as security, smart banking, marketing, crime investigation and in medical applications. Much research is also going in designing the robots which can detect the mental state of a patient by voice interaction mechanism. According to a report by WHO, number of suicidal attempts are also increasing year by year, so emotion recognition can also help in analyzing the mental status of patient having symptoms of anxiety and depression so that necessary preventive measures can be taken. This research paper reviews different proposed work done in the area of emotion recognition. Our work will be extended to developing such system for interaction with some AI based applications such as playing music based on one's emotions.

2. LITERATURE REVIEW

We have collected various research papers that deal with different aspects of emotion recognition system. Most of the time, results of these studies show that these systems can get good results by utilizing Machine Learning techniques such as Support Vector Machines and Deep Learning algorithms such as Convolutional Neural Network ,Artificial Neural Networks and Recurrent Neural Networks etc. With the help of MFCC and MLP classifiers CNN and SVM give a decent accuracy of around 80-90 %.

Bharti and Kukana proposed an approach for recognizing emotions using MSVM technique. The

authors divided the approach in four stages: Collecting the input, Extracting important features, Detection or Recognition and finally predicting the emotional state from Sad, Happy and Joy categories. The input samples were taken from the RAVDESS data set. They used Gammatone Frequency Cepstral Coefficient (GFCC) as feature extraction method. The work was simulated on MATLAB and the evaluation metrics used were AUC curve, False Alarm Rate, FRR, MSE, and SNR. This research work had concluded that the MSVM, GFCC, and ALO models have achieved better accuracy and SNR when compared with SVM and MFCC algorithms [1].

Another proposed research demonstrated by the authors in [2] focused on two different audio sets namely recorded audio and real time audio. The model includes four steps from signal acquisition to modeling of extracted features. The model focused on detecting continuous and qualitative speech. The continuous speech features carry linguistic contents of the emotional states and valuable information can be extracted from the glottal wave form which contains glottal excitation and vocal filters.

In the first step, input signal was taken in four different emotional states. The second step of Voice extraction was further divided into three parts- voice activity detection (to select significant portions of input signal), voiced portion accumulation (to achieve glottal pulse concentration) and data preprocessing using Linear Predictive Cepstral Coefficient for pitch detection. The next stage involves extracting features from the preprocessed data and eight features were extracted and in the final stage, five fold cross validation technique was used in classifying recorded data and implemented Glottal Pulse Feature technique with different classifiers on real time audio data. The results showed that proposed GPF model showed better accuracy in classifying certain emotional states . The strength of the proposed model is that the researchers have tried to design a technique that detects intense as well as mild expressing emotions accurately. The model efficiently finds multi-emotional states like fear, anger, happiness, disgust and sadness in English, Italian, Spanish multilingual platforms.

Authors of [3] proposed a model based on melfrequency cepstrum coefficient (MFCC).The objective of MFCC is to produce an output coefficient which works as input for the Hidden Markov Model to be classified into the speaker's emotions. The model was designed to identify seven certain types of emotions. The classification outputs are angry, calm, scared, happy, sad, disgusted and shocked. The data set was taken from SMART LAB from Ryerson University. These datasets were further divides into two parts for training and testing purposes.

The main feature of proposed methodology, Mel-frequency cepstrum, was used to extract sound signal characteristics based on the principle of hearing characteristics of the human ear. Then the testing is initiated through Hidden Markov Model and the categorization can be done. Evaluation was done on a dataset of 240 utterances and showed that the model gives an accuracy of 81.65% which is quite high [3].

A research work proposed in [4] demonstrated that modified mean cepstral features give improved recognition ratewhen compared to other conventional methods. The input data was taken from two speech databases EmoDB and SAVEE . The average recognition rates were 97.01% and 98.8% for EmoDB and SAVEE emotional speech datasets respectively. Another attempt in the field of emotion recognition is done in [5] by combining speech and facial features. The model was implemented in steps. These steps are as following:

i.Pre processing of the speech and facial data.

ii.Feature extraction using Deep Neural Network.

iii.DNNs for fusing emotional by sensing acoustic properties and features of face expression.

In the feature extraction face ID CNN and bi-directional log short term memory (Bi-LSTM) for extraction of temporal acoustic features was used. And for facial features extraction, multiple small scale kernel convolutional blocks were implemented.

The proposed model worked on the dataset IEMOCAP recorded by University of Southern California. The dataset contains data in terms video, audio, voice text, facial expressions. It is about of 12 hour. Detailed description of thedataset can be read from [11].

The main strength of the paper is that the proposed model not only showed a great improvement over uni-modal features model but also the use of multiple small scale convolution kernels

confirmed a good recognition rate along with minimized training parameters when compared to multiple large scale convolution kernels.

Authors of [6] proposed a Emotion Recognition System by creating two models. In the proposed system, feature extraction was done through MFCC. The first model was created using multi layer perceptron through Artificial Neural Network(ANN) and the second model was created using LSTM through Recurrent Neural Network(RNN).The dataset had 1440 audio records and was taken from the "Ryerson Audio Visual Database of Emotional Song and Speech". The system was implemented in Python with the libraries such as Tensorflow, Keras, Numpy, Matplotlib, Pandas, Librosa, Wave, Scikit-Learn. The model implementing MLP gave results with accuracy of 57.29% and the model implementing LSTM gave results with accuracy of 92.88%.

Research work in [7] proposed a model that aims at identifying emotion by facial expression. A facial expression plays an important role in identification of emotions. The model proposed in paper is initiated by taking real time video in input. It detects the face using local binary patterns cascade classifier.

The feature extraction step is performed after applying certain preprocessing on the data obtained. This model proposed an approach in which features are extracted by CNN with Histogram Oriented Gradients (HOG) and facial landmarks. CNN model having an input layer, four layers of convolution, two layers of pooling and two fully connected layers is the overall structure used for classification.

The dataset, used for classification was taken from two publicly available databases namely Japanese Female Facial Expressions (JAFFE) and FER2013. Both the datasets are popular in facial emotion recognition field. The. Experiment results on two databases shows that the proposed method can achieve an excellent performance. Accuracy of 91.2% and 74.4% was obtained on JAFFE and FER2013 database respectively.

In another work, an audio-text training paradigm was suggested by the researchers [8]. The Word2Vec and Speech2Vec models were trained, and the alignment of their two embedding spaces was done in such a way so that the Speech2Vec features were as similar to the Word2Vec features [10]. The convolution recurrent neural network is used to train the voice signal's low level characteristics. Before making the final prediction, a long short-term memory (LSTM) module that records the temporal dynamics in the signal is used to merge the semantic and paralinguistic characteristics into a single representation . The proposed work was simulated on the Sentiment Analysis in the Wild (SEWA) dataset, which was also used in the Audio/Visual Emotion Challenge (AVEC).

Another work [9] in the related field has been done while considering acoustic features of the speech. The proposed model worked on the combination of prosody, spectral and statistical functional feature values. The intention behind so is to enhance the progress of the classifier. For extracting useful information different statistical functions- mean, variance, range and standard deviation were taken into consideration to identify the emotional states. The proposed approach used KNN, LDA, and SVM techniques for classification. The extracted feature subset is taken as input to classifier. The classifier is trained using five proven speech corpus Emo-DB, SES, IITKGP-SESC, IEMOCAP, and IITKGP-SEHSC. These entire speech corpuses are benchmarked and contain variety of samples.

The results showed that SVM produces 93.5% of accuracy which is best among all three on IITKGP-SEHSC dataset. The strength of the paper is the use of feature fusion technique which was applied on relevant features and optimal feature selection to reduce high dimensionality problem . It enhanced the model's performance in terms of accuracy. But at the same time the obtained results show variance in the performance of a model for positive and negative emotions.

3. LIMITATIONS

Emotion recognition has become an important problem in Human Computer Interaction technology. Since last decade, much work has been done and in progress, but there are many challenges that come across during designing of robust and efficient SER systems. One such challenge is the background noise which occurs in real time input acquisition. Though the data sets

include audio with minimum or no noise but when it comes to taking real time data, noise is an obvious ingredient to the audio input. Another common challenge is the dimensional reduction and feature selection. Sometimes, feature selection comes with a trade –off of removing significant data from the input. Other problem occurs with the selection of the classifier based on the input and the features selected. We have seen that there are many conventional approaches which are in use in SER systems. When additional features are used with the classifiers, it outperforms the conventional methods.

4. CONCLUSION

In this paper, we have tried to put precise analysis of some emotion recognition systems. Such systems needed speech databases. These databases provide the data not only for the training process but also for testing purposes. However, the data needs certain preprocessing before applying in training or testing of model. Certainly preprocessing and relevant feature extraction plays an important role in the performance of the ML/DL techniques used. The validity and richness of data triggers the researchers to develop a better solution. The observed feature extraction techniques used in these papers are MFCC LPCC, DNN and LSTM. After the feature extraction, classification models are used to train the model and produce the required results. This paper focused on many of such techniques on which work has been done so far. This concludes that much work has been done in this field and when additional features such as semantics, acoustic , facial expression are combined with the audio data, and appropriate classifiers are used, results are much better than conventional classifiers.

5. REFERENCES

[1] Deepak Bharti, Poonam Kukana "A Hybrid Machine Learning Model for Emotion Recognition From Speech Signals" , Proceedings of the International Conference on Smart Electronics and Communication (ICOSEC 2020) IEEE Xplore Part Number: CFP20V90-ART; ISBN: 978-1- 7281-5461-9.

[2] Nazia Hossain, Mahmuda Naznin "Finding Emotion from Multi-Lingual Voice Data", 2020 IEEE 44th Annual Computers, Software, and Applications Conference (COMPSAC).

[3] Didik Muttaqin, Suyanto Suyanto, "Speech Emotion Detection Using Mel-frequency Cepstral Coefficient and Hidden Markov Model", 2020 3rd International Seminar on Research of Information Technology and Intelligent Systems (ISRITI) | 978-1-7281-8406-7/20/$31.00 ©2020 IEEE | DOI: 10.1109/ISRITI51436.2020.9315433.

[4] Krishna Chauhan, Kamlesh Kumar Sharma, Tarun Varma," Improved Speech Emotion Recognition Using Modified Mean Cepstral Features", 2020 IEEE 17th India Council International Conference (INDICON) | 978-1-7281- 6916- 3/20/$31.00 ©2020 IEEE | DOI: 10.1109/INDICON49873.2020.9342495.

[5] Linqin Cai, Jiangong Dong, Min Wei "Multi-Modal Emotion Recognition From Speech and Facial Expression Based on Deep Learning", 2020 Chinese Automation Congress (CAC) | 978-1-7281-7687-1/20/$31.00 ©2020 IEEE | DOI: 10.1109/CAC51589.2020.9327178.

[6] Shambhavi Sharma, " Emotion Recognition from Speech using Artificial Neural Networks and Recurrent Neural Networks", 11th International Conference on Cloud Computing, Data Science & Engineering (Confluence 2021)

[7] Dr Ansamma John , Abhishek MC , Ananthu S Ajayan, Sanoop S and Vishnu R Kumar, "Real-Time Facial Emotion Recognition System With Improved Preprocessing and Feature Extraction" , Proceedings of the Third International Conference on Smart Systems and Inventive Technology (ICSSIT 2020) IEEE Xplore Part Number: CFP20P17-ART;ISBN: 978-1-7281-5821-1;

[8] Panagiotis Tziraki , Anh Nguyen , Stefanos Zafeiriou , Bjorn W. Schuller, "Speech emotion recognition using semantic information", IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 978-1- 7281-7605- 5/20/$31.00 ©2021 IEEE | DOI: 10.1109/ICASSP39728.2021.9414866.

[9] Divya Lingampeta, Bhanusree Yalamanchili," Human Emotion Recognition using Acoustic Features with Optimized Feature Selection and Fusion Techniques", 2020 International Conference on Inventive Computation Technologies (ICICT)

[10] Y.-A. Chung, W.-H. Weng, S. Tong, and J. Glass,"Unsupervised cross-modal alignment of speech and text embedding spaces," in Proc. Advances in neural informationprocessing systems (NeurIPS), 2018, pp. 7354–7364

[11] C. Busso, M. Bulut, C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. Chang, S. Lee, and S. Narayanan, "IEMOCAP: Interactive emotional dyadic motion capture database," Journal of Language Resources and Evaluation, vol. 42, no. 4, pp. 335-359, December 2008.