Audio Analysis and Classification: A Review

Aas Mohammad, Dr. Manish Madhava Tripathi

Abstract— Communication plays a vital role according to the people's emotion, as emotions and gesture play 80% role while communication. Nowadays emotion recognition and classification are used in different areas to understand the human feelings like in the robotics, Health care, Military, Home automation, Hands-free computing, Mobile Telephony, Video game,call-center system, Marketing, etc. SER can help better interaction between the machine and the human. There are various algorithms and combination of the algorithms are available to recognize and classify the audio according to their emotion. In this paper, we attempted to investigate the episodic significant works, their technique and the impact of the approaches and the scope of the correction of the results.

Keywords:-emotion recognition, MFFC, Energy feature,

I. INTRODUCTION

Emotion is a very important subject in human being its play a very important role in peoples communication. In previous era communication between human, a machine is so difficult. Today interaction between human an machine is a very common thing. Training the computer to learn human emotion is an important facet of this communication. there are various applications and electronic devices are available for human emotion recognition in the market. these applications are available in car,cell-phone, computer, and televisions, etc. To develop a computer to learn human emotion and give a better communication experience is so challenging task. emotion recognition is the most popular research field.there are a general way to recognize the people's voice emotion is extracting the features that are respective to different emotional states to give these feature as input end of a classifier and get various emotion at the output end.

Here the main objective is to classify the audio emotion in four categories such as happy, sad, angry, natural. Prior to feature extraction pre-processing is applied to the audio.

Audio samples are contracted from voice and analog speech signals are change into the digital signals. After pre-processing every sentence is normalized to confirm that total sentences have the same volume range. Finally division of individual signal in the frames, therefore, speech signal keep up its features in short time.genrally used features are selected for the study after extracted. Emergy is an important fundamental feature of an audio signal.pitch is repeatedly used in this subject and autocorrelation is applied to identifying the pitch from every frame. After autocorrelation, statistical values are scheduled for speech signals.

Formant is a different crucial feature. for extraction of the first formant Linear Predictive Coding method is used. Later the autocorrelation statistical values are scheduled for the first formant. Mel frequency cepstral coefficient (MFCC) is an illustration of temporal energy spectrum on a people like mel scale frequency. Initial three coefficients of MFCCs are taken to infer means transformation. whole features of the audio samples are placed in the Artificial Neural Network.ANN comprises of an input matrix and with the objective matrix. which demonstrate the emotion state for each sentence made the input of the neural network. After this, we perform the classification procedure. Artificial neural network (ANN)used the training and test dataset for classification.

II. APPROACH

Emotion recognition has included the analysis of people expression has a different form like text, audio, videos.peoples have detected their emotion such as facial expression, body movement gesture, and speech.there are three main way to classify the emotion knowledge-based techniques, statistical methods, and hybrid approaches.

Knowledge-based techniques:- Knowledge-based techniques are also called the lexicon-based technique. In this approach, during the emotion classification process commonly used the knowledge-based resources for the WordNet, SenticNet, ConceptNet, and EmotiNet, the pros of this approach is openness and economy realized by the substantial accessibility of such knowledge-based resources. the cons of this technique are its failure to deal with idea subtleties and complex linguistic rules. The knowledge-based technique is mostly divided into two classifications dictionary-based method and corpus-based methodologies. In the dictionary-based approach is to identify the emotional words in a dictionary and it is also finding their antonyms and synonyms to enlarge the basic list of emotion.on another hand corpus-based approach begins with a seed list of emotional words an enlarge the database by searching different word with context-specific features in a huge corpus.

Manuscript revised June 9, 2019 and published on July 10, 2019

Aas Mohammad, *M.Tech Degree in Computer Science & Engineering at Integral University, Lucknow (U.P.).*
Dr. Manish Madhava Tripathi, *Associate Professor in the Department of Computer Science & Engineering at Integral University ,Lucknow (U.P), India.*

doi: 10.32622/ijrat.76201926

Statistical methods:- Statistical method is generally used in various supervised machine learning algorithms in this a huge set of define data is fed into the algorithms for a method to learn and anticipate the proper emotion types. In this methodology of have two types of dataset. Training dataset and test dataset. where the previous is utilized to get learn with the qualities of the data. while the last is used to approve the performance of the machine learning algorithm. Machine learning computation is mostly conferred more proper categorization accuracy in comparison to different methodologies. for the good results in the classification, the procedure has a requirement to adequately spacious training set. Generally, machine learning computation involves Support Vector Machines (SVM), Naive Bayes, and Maximum Entropy. Deep learning, is a part of the unsupervised clan of ML.this is broadly used in the emotion recognition.there are various deep learning algorithms like ANN, CNN, LSTM, and ELM. the deep learning methodology is in the area of emotion recognition might be primarily credited to its achievement in a related application, for example, computer vision, speech recognition, NLP.

Hybrid approaches:- Hybrid approaches is a mixture of the Knowledge-based techniques and Statistical method which utilize supplementary qualities from twain approaches. A part of the works that have applied a group of knowledge-driven linguistic elements. Statistical methods incorporate sentic computing and iFeel, both of which have received the idea level Knowledge-based asset SenticNet. The contribution of knowledge-based assets in the usage of the hybrid methodology is most vital in the emotion classification procedure. Since Hybrid procedures gain from the advantages offered by both knowledge-based and statistical methodologies.they will, in general, have better classification performance instead of utilizing or knowledge-based or statistical methods independently. A drawback of utilizing hybrid techniques, be that as it may, is the computational complexity during classification procedure.

III. LITERATURE REVIEW

Li Zheng [1] in 2018 Proposed method is a combination of CNN and random forest. CNN is applied for feature extraction from speech emotion. The outcomes of the experiment CNN-RF models are better than the CNN model. This model is used in the Nao robot. the original record sound box is improved and another Recorder box is achieved. Recorder box not exclusively can meet format prerequisites of speech emotion recognition yet additionally can address issues of studying NAO- BASED speech signal investigation Chinese words division and speech recognition. After an efficient test, the proposed CNN-RF model gives NAO robot fundamental elements of speech emotion recognition.

Kyoungju Noh [2] in 2018 proposes a method of speech emotion recognition (SER) system utilizing client self-referential speech characteristics. Which follow the ROC (Rate of change) of the relating characteristics an incentive as according to the client's emotional states. the results demonstrate that extra self-referential methodology could be the solution for the cold-begin issue with little datasets which comprises speech recording of every individual expressed with a few emotions.

Nikolaos Vryzas [3] in 2018 The author concentrates on the analysis of emotional states differentiation potentials, in the adaptive approach goal at the making of a powerful multimodel speech emotion recognition system. In this experiment, a different classification algorithm is used to compare to the outcomes of a generalized/ augmented to multiple speakers emotional speech database. speech based application is developed for real-time sentiment analysis. It has two modules are combined audio recording tools and camera and speech-to-Text modules. and this system has user-friendly GUI, and audio, video, and text-files are stored growth of the personalized database

Panagiotis Tzirakis [4] in 2018 The author proposed a novel model for sustained speech emotion recognition. this method is instructed end-to-end, it has also included the CNN.this is utilized to extract the characteristics from voice. and stacked over it a 2-layer LSTM. Besides, we demonstrate the connection among Kernal and pooling size of the 1-d layers of our model, and window and step estimate for conventional sound highlight like MFCCs.

K. Tarunika [5] in 2018 proposed the idea to used Deep Neural Network (DNN) and k-nearest neighbor (k-NN) in speech emotion recognition particular spooky mind.in this research principle firm applications in palliative consideration. Under most exact result the alert signals are made through the cloud. Statement level feature extraction emotion classification.speech and emotion detection are the most powerful things in the upcoming instrument and system.

Saurabh Sahu[6] in 2018 author purpose a method to test with the application of antagonistic training method to increase the accuracy of a deep neural network (DNN) of speech emotion recognition using speech signal.here two training method are presented.first is adversarial training and second is virtual adversarial training in adversarial training for the training data given labels are based on the adversarial direction.in the virtual adversarial training output distribution for training, data is based only on the adversarial side. To show the efficiency of adversarial training method by applying the K-fold cross-validation test on the Interactive Emotional Dyadic Motion Capture (IEMOCAP) and a cross-corpus exhibition investigation on three distinct corpora. Outcomes determine reformation over a purely supervised method..three evolutions are using for cross-corpus setup and one corpus evolution on the IEMOCAP dataset. both cases we watch an improvement in the classification execution using the adversarial techniques.

Boris Puterka [7] in 2018 author proposed a method for speech emotion recognition to analysis the time of the results. In the SER used the CNN and spectrograms used as feature extraction. In the analysis, two convolutional layers are followed by the pooling layer and the other one fully-connected layer followed by the dropout and softmax layer on the output. In this paper aim to search the connection between the time of speech signal and seven emotion recognition rate.

doi: 10.32622/ijrat.76201926

Turgut Özseven [8] in 2018 In this paper, speech emotion recognition are processing speech with DSP approach to gain the emotional expository feature and the old approach are used in this process is acoustic analysis. In this paper, we are analyzing the impact frame size used in the framing in emotion recognition.according to the outcome gaining frame size is different according to the used dataset. In the emotion the recognition if 1ms is changed in frame size then effect the rate of change in emotion recognition.

Po-Wei Hsiao [9] in 2018 the author introduces a method to learn the most powerful or instructive division in the input signal. deep recurrent neural network method obtains 37.0% unweighted averaged recall rate, which is based on the official Hidden Markov Model (HMM) baseline method for dynamic modeling structure. the introduce attention mechanism on the braid baseline deep recurrent neural network method obtain 46.3% unweighted averaged recall rate. This is the best unweighted averaged recall rate ever obtained on FAU-Aibo practice in the dynamic modeling structure.

Alexander Iliev [10] 2018 the author introduces a method to extract sentiment recognition from speech signals that can apply to any media respective service.there has proposed a method for searching finding and advising digital media text-based pre-set of metadata content queries sorted out in two parts at that point of mapped to speech sentiment prompts extracted from the emotion layer of speech alone We additionally represent the difference in sentiment expression for male and female speakers and further propose that this separation may improve system performance.

Pavol Harár [11] in 2017 The author introduces a methodology for Voice emotion recognition with DNN tectonics utilizing convolutional, pooling and fully attach flakes. In this method, we used the German corpus (Berlin database for speech emotion recognition) int this we used 3 types of emotion like anger, natural, sad this database hold the 271 labeled recording that has a total length 783 seconds. All file break into the 20-millisecond division without overlap. Speech activity identification approach utilized to remove the silent segment and shatter all data into Train(80%) and Validation(10%) and Test(10%) sets. the deep neural network used the Stochastic Gradient Descent. For input, we are recipient raw data without feature selection. This module has an accuracy of 96.97% on all file classification.

Mohan G hai [12] in 2017 The main objective of this paper to speech emotion recognition and categorized into seven emotional states like anger, boredom, disgust, anxiety, happiness, sadness and neutral. The given method is based on the Mel Frequency Cepstral Coefficients (MFCC) used the Berlin database of emotional speech. Speech feature is extracted and converted into the feature vector. For classification, many different algorithms are used like Support Vector Machine, Random Decision Forest and Gradient Boosting. The random forest has predicted the correct emotion and give the highest accuracy of 81.5%.

Margarita Kotti [13] in 2017 The author introduces a method to recognize the emotion from movies and drama clips.focus on the emotion to differ between the angry, happy, and neutral. we extract the feature and subset which is not mostly used in the emotion recognition task. and used the support vector machines and random forest algorithm for feature extraction and classification and accuracy are obtain to 65.63%.when we use the K-nearest neighbor(KNN) classifier the accuracy have to obtain 56.88%. the exploited feature is verified and classification schema is applied on the feature set. We increase the large set of 1582 feature then we have obtained the accuracy 61.25%.

(Amol S Patwardhan [14] in 2017) Describe a method that automatically finds emotion multimodel audio and video data. In video data to search the emotion to identify the facial manifestation, head, hand pointing, and body agitation. Spectral features and prosodic features are extracted from audio.using the infrared sensor (Kinect) to record the deep information from audio and video files. OpenEar toolkit and Face API are used to analyzing the features. support vector machine and combined feature vector based classifier are used to the emotion detection. This model has overall accuracy 96.6%.it has also different module accuracy for facial expression(92.4%) and head movement (94.3%) and hand gesture has (77.5%) and body movement (65.2%)alone.

(Danqing Luo [15] in 2017) Describes a method of an ensemble speech emotion recognition using the Recurrent neural network and WRN as a base classifier. In this, we use the two classifiers and to improve accuracy. The recurrent neural network used to predicts the statement level and WRN predict the segment level and learn a representation of spectrogram.RNN has an Arrear display for voice emotion recognition task. Residual network (ResNet) in image related classification.

doi: 10.32622/ijrat.76201926

IV. COMPARATIVE STUDY

S. NO	STUDY OF PAPER	PROS	CONS	RESULT
[1]	CNN is used for feature extraction from speech emotion. random forest is utilized for assortment to classify the speech emotion characteristic.	In this method, the original record sound box is improved and another Recorder box is achieved.	In this surprise and fear of emotion to easily to confused and "happy" have lower than the average level.	The natural emotion has the highest recognition degree it has an accuracy of 90%.
[2]	Input is combined with user self-referential characteristics that return the rate of change of voice characteristics values according to the user's emotional state.	In this, the angry and happy emotion has energy values are higher than to sad and natural emotion states.	In this natural and sad emotion state have low accuracy.	Accuracy depends on the combination of input speech characteristics. Speech emotion recognition results show the recognition rate of test data is higher than cross-validation.
[3]	concentrate on the analysis of emotional states differentiation potentials, in the adaptive approach goal at the making of a powerful multimodel speech emotion recognition system.	The user can record the emotion speech and play these files and edit these files.	It has supported five class of emotion states like anger, disgust, fear, happiness, sadness, it did not support the natural and calm emotion states.	Here they represent the five class of emotional states outputs.that are based on the classification. The accuracy of this method is good.
[4]	this method is instructed end-to-end, it has also included the CNN.this is utilized to extract the characteristics from voice. and stacked over it a 2-layer LSTM.	In this study has categorized into the gender age and it has also focused the mother tongue.	This study focused only on some languages like French, Italian, German, Portuguese.	This model can easily predict the speech emotion, and videos valence dimension. this model has a better display to other traditional models.
[5]	Deep Neural Network (DNN) and k-nearest neighbor (k-NN) in speech emotion recognition and Statement level feature extraction emotion classification.	In this method checked the mood of people.this is also worked on both audio and video files.	To know the separate accuracy of audio and video then the video has low accuracy and audio have some better to the audio.	In the result, audio and videos have separate accuracy but we combine both it has given good accuracy.it have overall accuracy 95%.
[6]	the author uses two sets of assessment single corpus assessment on the IEMOCAP dataset and 3 assessments involving the cross-corpus setup. In both classification performance are improved using the adversarial methods.	The model which is trained by the adversarial training method has better performance to the baseline DNN.	Baseline DNN has a particular "HAPPY" samples have a low classification performance.	Adversarial training applies the smoothness of the output probabilities in this we use the IEMOCAP dataset for single and three evolutions. both classification performances are improved using the adversarial methods.

doi: 10.32622/ijrat.76201926

[7]	In the speech emotion, recognition using the CNN and spectrograms used as feature extraction.	When we increase the length of analyzing speech the method accuracy has been increased.	Analyze Speech is short then its have low accuracy and low performance.	In this approach has acquired better outcomes on voice emotion. method accuracy is based on the size of the analyze speech.
[8]	Speech emotion recognition is processing speech with DSP approach to gain the emotional expository feature and the old approach are used in this process is acoustic analysis.	the frame size is size change 1ms effect noticeable changes in speech emotion recognition.	It has analyzes is used only 4 types of dataset.and the length of audio recording is neglected.	As per the outcomes, the most reasonable frame sizes for EMOVO, EMODB, SAVEE, and eNTERFACE05 are 25ms, 30ms, 40ms, 26ms, individually.
[9]	deep recurrent neural network method obtains 37.0% unweighted averaged recall rate, which is based on the official Hidden Markov Model (HMM) baseline method for dynamic modeling structure.	RNN which have 37.0%, unweighted accuracy and LSTM have 46.3% unweighted accuracy which is the best performance in a dynamic modeling system of FAU-Aibo assignment.	Rest class have to hard the recognize the data. comfort class holds all labels to not belonging to the other four class.	The method has good performance and LSTM have unweighted accuracy is good as comparison to the RNN and HMM.
[10]	the author introduces a method to extract sentiment recognition from speech signals that can apply to any media respective service and it also classifies the gender.	In this method analyzes the different emotional states with gender Male or female and it has better performance to the sad state of emotion.	In this happy emotional state have a lower performance to the sad state.	In this approach determine the different emotion of the speech and the overall performance of this method is good.
[11]	the author introduces a methodology for Voice emotion recognition with DNN tectonics utilizing convolutional, pooling and fully attach flakes.	predict the different emotional state of a person and accuracy is depends on the testing layers. If increase the testing segment the accuracy is improved.	In this paper, accuracy is based on the German data set and it works only the single dataset not the multiple datasets.	In this method obtained 97.97% accuracy on the testing data with the general assuredness of 69.55% on file prediction.
[12]	speech emotion recognition and categorized into seven emotional class. The given method is based on the Mel Frequency Cepstral Coefficients (MFCC) used the Berlin database of emotional speech.	Random Decision Forest, SVM, Gradient Boosting, In these the Random Decision Forest is best and in this anger, the class has the highest accuracy.	In this, happiness has the least accuracy.	In the three algorithms, Random Decision Forest has the best accuracy and it has overall 81.05% obtained.

doi: 10.32622/ijrat.76201926

[13]	Focus on the emotion to differ between the angry, happy, and neutral. we extract the feature and subset which is not commonly used in the emotion recognition task.	In this method make a classification committee to takes on audio file as input and gives the output according to the emotion.here to two classifiers are used that increase the performance.	The dataset sample is a small database is used to train an test the method. and a small amount of segment layer is used.	The method has an overall 65.63% accuracy.
[14]	In video data to search the emotion to identify the facial manifestation, head, hand pointing, and body agitation. Spectral features and prosodic features are extracted from audio.	Body and hand gesture, angry and disgust has higher accuracy as compared to the other class.the method performance is very good.	The cons of this method are some feature has good performance and some do not accurately work.	some module has good accuracy and some have fair but combined all modules it gives the 96.6% accuracy.
[15]	Ensemble speech emotion recognition using the RNN and WRN as a base classifier. In this, we use the two classifiers and to improve accuracy.	Angry and natural subsystem in WRN is better than RNN subsystem. But the whole system RNN can recognize better as compared to the imbalance data.	RNN based subsystem not to find accurately the angry and natural as compared to the WRN subsystem but it works accurately on the Sad class.	comparison to the RNN-based and WRN-based speech emotion recognition the proposed ensemble system gives the 2% and 3% improvement in unweighted accuracy and weighted accuracy.

So we find from the table that, audio analysis can be done by using methods, CNN, ANN, RNN, and LSTM. The benefits of using CNN is when a model is trained than the prediction is done really quickly and it does not depend on the number of input and layers for training. The drawback of the CNN is it depends on the dataset. It has needed a more large dataset. if give small dataset it gives the slow performance in the results.according to the different paper survey efficiency of CNN is good. The pros of using ANN have parallel Processing capability and ANN have the numerical strength that performs more than one at the same time.ANN learns to events and makes the decisions by commenting on similar events. The Cons of the ANN performance depends on the Hardware. **The** network is reduced to a certain value of the error on the sample means that the training has been completed. This value does not give us optimum results. The Advantages of the using RNN can deal with consecutive information of licentious length. In the RNN one too many, many to one and many to many input and output are presumable. The Disadvantage of the RNN can't be stacked into extremely deep models. It is entirely shaky if we use ReLu as its enactment work. The benefits to use of the LSTM is Advanced part of RNN. LSTM can by lapse keep the data for a long duration. It is utilized for handling, predicting and classifying on the base of the period of series information. the drawback of LSTM is slower than other ordinary activation functions, like sigmoid, tanh or rectified linear unit.

V. CONCLUSION

In this survey, we demonstrate different approaches of emotion recognition of the peoples. A lot of suspicions are present for the finest algorithm to classify the emotion.there are a various combination of algorithms are available that gives different results and accuracy. we analyze the different algorithms give different results sometimes it depends on the size and type of the dataset. we investigate the different researches with their merit and demerit. the use of CNN and LSTM that gives better performance yet with different disadvantages. This paper surveys significant features in the ongoing developments and research of speech emotion recognition various method gives the technological point of view on society.

REFERENCES

[1] Li Zheng, Qiao Li, Hua Ban, Shuhua Liu(2018) "Speech Emotion Recognition Based on Convolution Neural Network combined with Random Forest" in Chinese Control And Decision Conference (CCDC).

[2] Kyoungju Noh, Seungeun Chung, Jiyoun Lim, Gague Kim, and Hyuntae Jeong (2018)"Speech Emotion Recognition Framework based on User Self-referential Speech Features" in IEEE 7th Global Conference on Consumer Electronics (GCCE).

[3] Nikolaos Vryzas, Lazaros Vrysis, Rigas Kotsakis and Charalampos Dimoulas(2018) "Speech Emotion Recognition Adapted to Multimodal Semantic

doi: 10.32622/ijrat.76201926

Repositories” in 13th International Workshop on Semantic and Social Media Adaptation and Personalization (SMAP).

[4] Panagiotis Tzirakis, Jiehao Zhang, Bj¨orn W. Schuller(2018) “End-To-End Speech Emotion Recognition Using Deep Neural Networks” in IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP).

[5] K.Tarunika, R.B Pradeeba, P.Aruna (2018) “Applying Machine Learning Techniques for Speech Emotion Recognition” 9th International Conference on Computing, Communication, and Networking Technologies (ICCCNT).

[6] Saurabh Sahu, Rahul Gupta, Ganesh Sivaraman, Carol Espy-Wilson(2018) “Smoothing Model Predictions Using Adversarial Training Procedures For Speech Based Emotion Recognition” in IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP).

[7] Boris Puterka, Juraj Kacur(2018) “Time Window Analysis for Automatic Speech Emotion Recognition” in International Symposium ELMAR.

[8] Turgut Özseven(2018) “Evaluation of the Effect of Frame Size on Speech Emotion Recognition” in 2nd International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT).

[9] Po-Wei Hsiao and Chia-Ping Chen(2018) “Effective attention mechanism in dynamic models For speech emotion recognition” in IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP).

[10] Alexander Iliev Peter L. Stanchev (2018) “Information Retrieval and Recommendation using Emotion from Speech Signals” in IEEE Conference on Multimedia Information Processing and Retrieval.

[11] Pavol Harár, Radim Burget and Malay Kishore Dutta (2017) “Speech Emotion Recognition with Deep Learning” in 4th International Conference on Signal Processing and Integrated Networks (SPIN)

[12] Mohan Ghai, Shamit Lal, Shivam Duggal and Shrey Manik (2017)“Emotion Recognition On Speech Signals Using Machine Learning” in International Conference on Big Data Analytics and Computational Intelligence (ICBDAC).

[13] Margarita Kotti and Yannis Stylianou (2017) “Effective Emotion Recognition In Movie Audio Tracks” in IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP).

[14] Amol S Patwardhan (2017) “Multimodal Mixed Emotion Detection” in 2nd International Conference on Communication and Electronics Systems (ICCES).

[15] Danqing Luo, Yuexian Zou, Dongyan Huang (2017) “Speech Emotion Recognition via Ensembling Neural Networks” in Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)

AUTHORS PROFILE

Aas Mohammad is student of integral university. He achieved B.TECH Degree from Integral University, Lucknow (U.P.), and currently pursuing his M.Tech Degree in Computer Science & Engineering at Integral University, Lucknow (U.P.). His research in Audio Analysis and Classification using deep learning .

Manish Madhava Tripathi is currently working as Associate Professor in the Department of Computer Science & Engineering at Integral University ,Lucknow (U.P),India .He has completed his PhD in “Designing a Model To Improve Medical Image Watermarking” .He has over 1 years Industrial experience and 17 year experience in Academics. He has published more than 40 papers inreputed Journals and Conferences..He is life time member of “Computer Society of India” and member of IEEE.

doi: 10.32622/ijrat.76201926