A Review Paper on Speech Based Emotion Detection Using Deep Learning

Prof. Martina D'souza[1], Rohan Adhav[2], Shivam Dubey[3], Sachin Dwivedi[4]

[1]*Assistant Professor Dept. of Information Technology, Xavier Institute of Engineering, Mumbai, Maharashtra, India*
[2,3,4]*Student, Dept. of Information Technology, Xavier Institute of Engineering, Mumbai, Maharashtra, India*

Abstract - *Feature extraction is a very important part in speech emotion recognition, and in allusion to feature extraction in speech emotion recognition problems. As emotions play a vital role in communication, the detection and analysis of the same is of vital importance in today's digital world of remote communication. Due to the complexity of the task of emotion recognition, it is commonly utilized to extract emotions from speech signals. Many techniques are used to achieve this goal. Deep Learning techniques have been suggested as an alternative to traditional methods in speech-based emotion recognition. Include extraction could be a exceptionally critical portion in discourse feeling acknowledgement, and in mention to include extraction in discourse feeling acknowledgement issues. As feelings playa imperative part in communication, the location and examination of the same is of crucial significance in today's computerized world if inaccessible communication. In human Computer Interaction (HCI) region discourse feeling recognition is one of the well-known theme within the world. Numerous analysts are locked in, in creating systems to recognize diverse feelings from human discourse. This is done to create HCI and human interface more viable and create frameworks like people. This paper represents an overview of these techniques and discusses their limitations in terms of speech-based emotion recognition.*

Key Words: **Speech Emotion Recognition, Deep Learning, Deep Neural Network, Deep Boltzmann Machine, Recurrent Neural Network, Deep Belief Network, Convolutional Neural Network, etc**

1.INTRODUCTION

Emotion recognition has evolved from a niche technology to an integral component of Human Computer Interaction. Some examples of applications include speech recognition for call centers, vehicle driving systems, and medical applications. However, there are still many problems that need to be addressed in order to achieve the best possible results. Different models are used to classify different emotions in humans. Among these is the discrete emotional approach, which considers all the emotions as a group. It classifies them into various categories such as anger, boredom, fear, and happiness. The approach for emotion recognition is usually divided into two phases: the feature extraction phase and the features classification phase. During the former, researchers have focused on various features such as vocalization factors and excitation features. Due to the non-stationary nature of the speech signal, it is considered that speech recognition is not prone to linearization. Therefore, non-linear classifiers such as the Gaussian Mixture Model and the Hidden Markov Model are widely used for emotion recognition. Energy based features such as Linear Predictor Coefficients (LPC), Mel Energy-spectrum Dynamic Coefficients (MEDC), MelFrequency Cepstrum Coefficients (MFCC) and Perceptual Linear Prediction cepstrum coefficients (PLP) are often usedfor effective emotion recognition from speech. For emotion recognition, other classifiers such as K-Nearest Neighbor (KNN), Principal Component Analysis (PCA), and Decision trees are used. Deep Learning is a branch of machine learning that focuses on learning complex structures and features without requiring manual intervention. It has gained more attention due to its ability to detect features and minimize the complexity of the data. Convolutional Neural Networks and Deep Neural Networks are commonly used for image and video processing. However, they can also be commonly used for speech recognition. Due to the complexity of their learning tasks, RNNs and CNNs tend to require large storage capacities. However, with the help of deep learning, they can handle different types of input data.

2.LITERATURE SURVEY

2.1 Speech Enhancement Using Pitch Detection Approach for Noisy Environment, Rashmi M, Urmila S, Dr. V.M. Thakare, IJEST 2011

Acoustical jumble among preparing and testing stages debases extraordinarily discourse acknowledgment comesabout. This issue has constrained the advancement of real-world nonspecific applications, as testing conditions are exceedingly variation or indeed unusual amid the preparing prepare. Subsequently the foundation commotion must be removed from the loud discourse flag to extend the flag comprehensible and to decrease the audience weariness. Upgrade procedures connected, as pre-processing stages; to the frameworks surprisingly make strides acknowledgment comes about. In this paper, a novel approach is utilized to upgrade the seen quality of the speech signal when the added substance clamor cannot be straightforwardly controlled. Rather than controlling the foundation commotion, we propose to fortify the discourse flag so that it can be listened more clearly in boisterous environments. The subjective

assessment appears that the proposed strategy moves forward perceptual quality of discourse in different boisterous situations. The subjective assessment appears that the proposed strategy makes strides perceptual quality of discourse in different boisterous situations. As in a few cases speaking may be more helpful than writing, indeed for quick typists: numerous numerical images are lost from the console but can be effectively talked and recognized. In this manner, the proposed framework can be utilized in an application planned for numerical image acknowledgment (particularly images not accessible on the console) in schools.

2.2 Emotion Recognition of Affective Speech Based on Multiple Classifiers Using Acoustic-Prosodic Information and Semantic Labels (Extended Abstract), Chung-Hsien Wu, Wei-Bin Liang, ACII 2015

This work presents an approach to feeling recognition of full of feeling discourse based on different classifiers utilizing acoustic-prosodic data (AP) and semantic names (SLs). For AP-based acknowledgment, acoustic and prosodic highlights are extracted from the identified enthusiastic notable sections of the input discourse. Three sorts of models GMMs, SVMs, and MLPs are embraced as the base-level classifiers. A Meta Choice Tree (MDT) is at that point utilized for classifier combination to get the APbased feeling acknowledgment certainty. For SL-based acknowledgment, semantic names are utilized to consequently extricate Feeling Association Rules (EARs) from the recognized word arrangement of the emotional discourse. The most extreme entropy show (MaxEnt) is thereafter utilized to characterize the relationship between emotional states and EARs for feeling acknowledgment. At long last, a weighted item combination strategy is utilized to coordinated the AP-based and SL-based acknowledgment comes about for last feeling decision. For assessment, 2,033 expressions for four enthusiastic states were collected. The test comes about uncover that the emotion acknowledgment execution for AP-based acknowledgment utilizing MDT accomplished 80.00%. On the other hand, an normal recognition exactness of 80.92% was gotten for SL-based recognition. At last, combining AP data and SLs achieved 83.55% precision for feeling acknowledgment.

2.3 Speech based Emotion Recognition using Machine Learning, Girija D, Apurva G, Gauri G, Sukanya K, ICCMC 2019

Emotion acknowledgment from sound flag requires feature extraction and classifier preparing. The highlight vector consists of components of the sound flag which characterise speaker particular highlights such as tone, pitch, vitality, which is crucial to prepare the classifier show to perceive a specific emotion precisely. The opensource dataset for North American English dialects was physically separated into preparing and testing. Speaker vocal tract data, spoken to by Mel-frequency cepstral coefficients (MFCC), was extricated from the sound samples in preparing dataset. Pitch, Brief Term Energy(STE), and MFCC coefficients of sound tests in feelings outrage, happiness, and pity were gotten. These extricated highlight vectors were sent to the classifier show. The test dataset will undergo the extraction method taking after which the classifier would make a choice with respect to the fundamental feeling within the test sound. The preparing and test databases utilized were North American English acted and characteristic discourse corpus, real-time input English discourse, territorial dialect databases in Hindi and Marathi. The paper points of interest the two strategies connected on include vectors and the impact of expanding the number of include vectors fed to the classifier. It gives an examination of the precision of classification for Indian English discourse and discourse in Hindi and Marathi. The accomplished exactness for Indian English discourse was 80 percent.

2.4 Speech Emotion Recognition Using Deep Neural Network considering Verbal and Nonverbal Speech Sounds, Kun-Yi Huang, Chung-Hsien Wu, Qian-Bei Hong, Ming-Hsiang Su and Yi-Hsuan Chen, ICASSP 2019

Discourse feeling acknowledgment is getting to be progressively important for numerous applications. In real-life communication, non-verbal sounds inside an articulation too play an imperative role for individuals to recognize feeling. In current ponders, as it were few feeling acknowledgment frameworks considered nonverbal sounds, such as giggling, cries or other feeling interjection, which actually exists in our everyday discussion. In this work, both verbal and nonverbal sounds inside an expression were thus considered for feeling acknowledgment of real-life conversations. Firstly, an SVM-based verbal/nonverbal sound locator was created. A Prosodic Express (PPh) auto-tagger was advance utilized to extricate the verbal/nonverbal fragments. For each fragment, the emotion and sound highlights were separately extricated based on convolutional neural systems (CNNs) and after that concatenated to create a CNN-based bland include vector. At last, a arrangement of CNN-based highlight vectors for an entire exchange turn was bolstered to an mindful long short-term memory (LSTM)-based sequence-to-sequence show to output an passionate grouping as acknowledgment result. The experimental is based on the recognition of seven ecstatic phases inside the NNIME (The NTHU-NTUA Chinese interactive multimodal feeling corpus) appeared that the proposed strategy accomplished a location exactness of 52.00% outperforming the conventional strategies.

2.5 Emotion Recognition from Speech based on Relevant Feature and Majority Voting, Md. Kamruzzaman S, Kazi Md. Rokibul Alam, Md. Arifuzzaman, IEV 2014

This paper proposes an approach to distinguish feeling from human discourse utilizing lion's share voting method over several machine learning procedures. The commitment of

this work is in two folds: firstly it chooses those highlights of discourse which is most promising for classification and besides it employments the lion's share voting method that chooses the precise course of emotion. Here, lion's share voting strategy has been connected over Neural Organize (NN), Choice Tree (DT), Bolster Vector Machine (SVM) and K-Nearest Neighbor (KNN). Input vector of NN, DT, SVM and KNN comprises of different acoustic and prosodic highlights like Pitch, Mel-Frequency Cepstral coefficients etc. From discourse flag numerous highlight have been extricated and as it were promising highlights have been selected. To consider a include as promising, Quick Relationship based highlight choice (FCBF) and Fisher score calculations have been utilized and as it were those highlights are chosen which are highly positioned by both of them. The proposed approach has been tried on Berlin dataset of passionate discourse and Electromagnetic Articulography (EMA) dataset. The experimental result appears that lion's share voting method attains superior exactness over person machine learning techniques. The work of the proposed approach can effectively recognize the feeling of human creatures in case of social robot, brilliantly chat client, call-center of a company etc.

2.6 Speech Based Human Emotion Recognition Using MFCC, M.S. Likitha, Sri Raksha R. Gupta, K. Hasitha and A. Upendra Raju, WiSPNET 2017

Discourse could be a complex flag comprising of various information, such as data approximately the message to be communicated, speaker, dialect, locale, feelings etc. Speech Processing is one of the imperative branches of computerized signal processing and finds applications in Human computer interfaces, Media transmission, Assistive advances, Sound mining, Security and so on. Discourse feeling acknowledgment is vital to have a normal interaction between human being and machine. In discourse feeling acknowledgment, passionate state of a speaker is extracted from his or her discourse. The acoustic characteristic of the discourse flag is Highlight. Include extraction is the process that extricates a little sum of information from the discourse signal that can afterward be utilized to speak to each speaker. Numerous feature extraction strategies are accessible and Mel Recurrence Cepstral Coefficient (MFCC) is the commonly utilized strategy. In this paper, speaker feelings are recognized utilizing the information extricated from the speaker voice flag. Mel Recurrence Cepstral Coefficient (MFCC) method is utilized to recognize feeling of a speaker from their voice. The planned framework was approved for happy, sad and outrage feelings and the proficiency was found to be about 80%.

2.7 Speech Emotion Recognition Using Support Vector Machine, Yashpalsing Chavan, M.L. Dhore, Pallavi Yesaware, IJCA 2010

Programmed Discourse Feeling Acknowledgment (SER) may be a current research subject within the field of Human Computer Interaction (HCI) with wide run of applications. The discourse highlights such as, Mel Recurrence cepstrum coefficients (MFCC) and Mel Vitality Spectrum Energetic Coefficients (MEDC) are extricated from speech expression. The Bolster Vector Machine (SVM) is utilized as classifier to classify diverse enthusiastic states such as outrage, happiness, pity, unbiased, fear, from Berlin enthusiastic database. The LIBSVM is utilized for classification of feelings. It gives 93.75% classification precision for Sex autonomous case 94.73% for male and 100% for female speech.

2.8 Emotion Recognition from Speech Using MFCC and DWT for Security System, Sonali T. Saste, Prof. S.M. Jagdale, ICECA 2017

In later a long time the feeling acknowledgment from speech is zone of more intrigued in human computer interaction. There are numerous diverse analysts which worked on emotion acknowledgment from discourse with distinctive frameworks. This paper endeavors feeling acknowledgment from discourse which is language free. The enthusiastic discourse tests database is utilized for include extraction. For include extraction MFCC and DWT these two distinctive calculations are utilized. For classification of diverse feelings like irate, upbeat, frightened and unbiased state SVM classifier is utilized. The classification is based on the highlight vector shaped by combination of two algorithms. This classified feeling is utilized for ATM security system.

2.9 Speech Emotion Recognition Using Convolutional Neural Network (CNN), Apoorv Singh, Kshitij Kumar Srivastava, Harini Murugan, IJPR 2020

The Mechanized Discourse Feeling Acknowledgment could be a extreme prepare since of the hole among acoustic characteristics and human feelings, which depends unequivocally on the discriminative acoustic characteristics extracted for a given acknowledgment assignment. Distinctive people have distinctive feelings and through and through a diverse way to express it. Discourse feeling do have diverse energies, pitch varieties are emphasized in the event that considering different subjects. Therefore, the discourse feeling location could be a requesting assignment in computing vision. Here, the discourse feeling recognition is based on the Convolutional Neural Organize (CNN) algorithm which employments distinctive modules for the emotion acknowledgment and the classifiers are utilized to distinguish feelings such as bliss, astonish, outrage, unbiased state, pity, etc. The dataset for the discourse feeling acknowledgment framework is the discourse tests and the characteristics are extricated from these discourse tests utilizing LIBROSA bundle. The classification execution is based on extricated characteristics. At long last we will decide the feeling of discourse signal.

2.10 Deep learning based Affective Model for Speech Emotion Recognition, Xi Zhou, Junqi Guo, Rongfang Bie, UIC-ATC 2016

Considering the application esteem of feeling, increasing consideration has been pulled in on feeling acknowledgment over the final decades. We commit ourselves to doable discourse emotion recognition investigate. We construct two full of feeling models based on two profound learning strategies (a stacked autoencoder arrange and a profound conviction arrange) separately for programmed feeling feature extraction and feeling states classification. The experiments are based on a well-known German Berlin Enthusiastic Speech Database, and the acknowledgment precision comes to 65% within the best case. In expansion, we approve the impact of distinctive speakers and distinctive feeling categories on acknowledgment exactness.

3. CONCLUSION

This paper has given a point by point audit of the profound learning strategies for SER. Profound learning strategies such as DBM, RNN, DBN, CNN, and AE have been the subject of much inquire about in later a long time. These profound learning methods and their layer-wise structures are briefly explained based on the classification of different common feeling such as happiness, bliss, pity, impartial, shock, boredom, disgust, fear, and outrage. These strategies offer simple show training as well as the proficiency of shared weights. Confinements of deep learning methods incorporate their huge layer-wise inner design, less effectiveness for temporally-varying input data and over-learning amid memorization of layer-wise information. This investigates work shapes a base to evaluate the execution and confinements of current profound learning techniques. Advance, it highlights a few promising directions for way better SER frameworks.

REFERENCES

[1] Speech Enhancement Using Pitch Detection Approach for Noisy Environment, Rashmi M, Urmila S, Dr. V.M. Thakare, IJEST 2011, ISSN: 0975-5462

[2] Emotion Recognition of Affective Speech Based on Multiple Classifiers Using Acoustic-Prosodic Information and Semantic Labels (Extended Abstract), Chung-Hsien Wu, Wei-Bin Liang, ACII 2015,978-1-4799-9953-8/15/$31.00 ©2015 IEEE

[3] Speech based Emotion Recognition using Machine Learning, Girija D, Apurva G, Gauri G, Sukanya K, ICCMC, 978-1-5386-7808-4/19/$31.00 ©2019 IEEE

[4] Speech Emotion Recognition Using Deep Neural Network considering Verbal and Nonverbal Speech Sounds, Kun-Yi Huang, Chung-Hsien Wu, Qian-Bei Hong, Ming-Hsiang Su and Yi-Hsuan Chen, ICASSP, 978-1-5386-4658-8/18/$31.00 ©2019 IEEE

[5] Emotion Recognition from Speech based on Relevant Feature and Majority Voting, Md. Kamruzzaman S, Kazi Md. Rokibul Alam,Md. Arifuzzaman, ELECTRONICS & VISION 2014 978-1-4799-5180-2/14/$31.00 ©2014 IEEE

[6] Speech Based Human Emotion Recognition Using MFCC, M.S. Likitha, Sri Raksha R. Gupta, K. Hasitha and A. Upendra Raju, WiSPNET, 978-1-5090-4442-9/17/$31.00 ©2017 IEEE

[7] Speech Emotion Recognition Using Support Vector Machine, Yashpalsing Chavan, M.L. Dhore, Pallavi Yesaware, IJCA, ©2010 International Journal of Computer Applications (0975 - 8887) Volume 1 – No. 20

[8] Emotion Recognition from Speech Using MFCC and DWT for Security System, Sonali T. Saste, Prof. S.M. Jagdale, ICECA, 978-1-5090-5686-6/17/$31.00 ©2017 IEEE

[9] Speech Emotion Recognition Using Convolutional Neural Network (CNN), Apoorv Singh, Kshitij Kumar Srivastava, Harini Murugan, International Journal of Psychosocial Rehabilitation, Vol. 24, Issue 08, 2020 ISSN: 1475-7192

[10] Deep learning based Affective Model for Speech Emotion Recognition, Xi Zhou, Junqi Guo, Rongfang Bie, 978-1-5090-2771-2/16 $31.00 © 2016 IEEE DOI 10.1109/UIC-ATC-ScalCom-CBDCom-IoP-SmartWorld.2016.42