Speech Emotion Recognition: A survey

Kunal Bhapkar[1], Krati Patni[2], Praddyumn Wadekar[3], Shweta Pal[4], Dr. Rubeena A. Khan[5], Mahesh Shinde[6]

---***---

Abstract - *The chief point of this paper is to supply an outline of Speech Emotion Recognition. Emotions can be recognized by extracting many features from the speech. In SERs, numerous methods have been resorted to remove sentiments from waves, tallying various well-established talk examination and classification methodologies. Recent work on emotion recognition using speech is carried out, and various issues related to this have been presented in this paper. The main challenges for recognizing an emotion are the selection of database, identification of various speech related features and an adequate choice of model for classification. We have done literature study on various characteristics used to recognize emotions from human speech. In addition to some recent study work reviews, the significance of different classification models have been presented.*

***Key Words*: Speech Emotion Recognition; Convolutional Neural Networks; Classification; Support Vector Machine; Mel frequency cepstral coefficients**

I. INTRODUCTION

Many features of the human vocal system such as speech, tone, pitch and many others, convey information and context [1]. For natural human–computer communication demands, SER is widely used. Speech emotion recognition, is noted as removing the passionate form of a speaker from his or her talk. This sort of recognition is supposed to be used to extract useful semantics of speech recognition systems.

The model of SER contains two types of models; the discrete speech emotion model and continuous speech emotion model [2]. The first model expresses several individualistic emotions, indicating that a certain voice has a single individualistic emotion, while the second one means that the emotion is in the emotion space, and every emotion possess unique strength on each proportion.

Concluding the emotional state of humans is an idiosyncratic task and can be employed as a level for any feeling recognition model. It uses diverse emotion such as disgust, anger, fear, surprise, joy, happiness, sadness and neutral.

The approach for SER primarily comprises of three phases known as pre-processing, feature extraction and feature classification phase [3].

- Pre-processing: Pre-processing applies to all the raw data transformations before it is fed into the machine. It involves the elimination of silence, pre-emphasis, normalization and windowing, so it is an essential step to get the pure signal used in the next stage i.e., in feature extraction. It is also important in order to speed up training.

- Feature Extraction: Without disturbing the properties of the speech, for examining the signal, a minor quantity of information from the speech signal is withdrawn [4]. Mel cepstral coefficients are frequently used feature parameters for speech recognition [5]. Based on the responsiveness of the hearing organ, MFCC uses the Mel scale [6]. In this study, from the speech signals some features are extracted and on this extracted features analyses are carried out [7]. Feature extraction requires multiple layers of convolution accompanied by max-pooling and an activation function. Using the various feature extraction algorithms, the speech emotion recognition rate of a device is increased. The work emphasizes the pre-processing of the audio samples obtained, where the noise is eliminated using filters from speech samples.

- Feature Classifier: For any pattern recognition in Speech Emotion Recognition mainly classifier can be divided into types, namely non-linear classifiers and linear classifiers [8].

There are various classification methods used to create the correct classifier to model emotional states. Such as Hidden Markov Models (HMM), SVM (Support Vector Machine), Gaussian Mixture Models (GMM), Neural Network and K-Nearest Neighbor.

II. LITERATURE SURVEY

Girija Deshmukh et al. in [1] proposed a system in which they obtained audio samples of Short-Term Energy (STE), Pitch, and MFCC coefficients in frustration, happiness, and sadness of emotions. Open source North American English served as expression and as feedback was used to record natural speech. Thus, only three emotions i.e., anger, happiness and sadness were recognized. They also identified the speaker's detailed features, such as sound, energy, pitch. The whole Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) dataset is manually split into train and test sets. The multi-class Support vector machine (SVM) takes feature vectors as input, which is turned up as a model corresponding to each emotion.

Peng Shi in [2] introduced discrete model and continuous model of speech emotion recognition; different characteristics are analysed to make better description of emotions. When compared to Artificial Neural Networks (ANNs) and support vector machines (SVMs), the Deep Belief

Networks (DBNs) have about 5% higher accuracy rate than the traditional methods. The output shows that the features which are extracted by Deep Belief Networks is much better than the original feature. DBN-SVM had slightly improved result than DBN-DNN because SVM classifies in small size better. DBN converts empty characteristics into deep abstract characteristics, resulting into better classification.

J. Umamaheswari et al. in [3] presented pre-processing being carried out using K-Nearest Neighbour (KNN) and Pattern Recognition Neural Network (PRNN) algorithms while the feature extraction was explained using a descended structure covering Gray Level Co-occurrence Matrix (GLCM) and Mel Frequency Cepstral Coefficient (MFCC). The outcomes were compared for their precision rate, accuracy and f-Measure with standard algorithms like Hidden Markov Model (HMM) and Gaussian mixture models (GMMs) were recognized as a better production than the benchmark algorithms. The emotional waves generate a pattern, the pattern of the signal is later recognized by PRNN. The believable nearest pattern concerning the signal is determined by K-NN approach. The Speech database contains six fundamental classes such as:

- Neutral
- Fear
- Angry
- Surprise
- Sad
- Happy

M.S. Likitha et al. in [4] observed recognition requires assessment of the verbal communication wave to classify the required feeling, based on the training of its characteristics, like Sound, format, phoneme. On the side of withdrawal of functionality and examination, A good number of algorithms were made of a speech signal. The acoustic precision of the communication kinesics is a feature. Withdrawal of features is the process of removing a compact amount of information from the voice signal employed to reflect each speaker later on. Most Methods of extraction are at one's fingertips but the widely used method is coefficient (MFCC). Feature is the sound representative of the speech signal. An action that mines a little quantity of information from the verbal expression that can subsequently be used to act for each speaker is called feature extraction.

Zhang Lin et al. in [5] conveyed that SER innovation is used to monitor the driver's irregular emotions, and makes use of particular phrase recognition innovation to pick out the parking guidance in emergency situations. Three types of speech features are extracted that are prosodic, spectral-based and quality features. Frequently used characteristic parameters in speech recognition are the Mel-cepstral coefficient (MFCC) and the Linear Predictive Cepstral Model (LPCC). SVM is used for extracting the features. The concern of panic in the driver's voice can be defined by way of this system as an emergency situation. The parking instructions are listed on that basis. Since the voice parking guidance database and the speech emotion database used in this paper are collected on various settings, the recognition efficiency is significantly degraded as soon as the emotion of the parking guidance voice is declared.

Asaf Varol et al. in [6] expressed how sound is defined as a pressure wave that arises from the vibration of substance's present in a molecule. The investigation had gone through sound energy and its characteristics. More effective results are drawn from experiments using the speech signal spectrogram and Artificial Neural Networks (ANNs). Using EMO-DB dataset, SER uses role extraction techniques such as acoustic analysis and analysis of spectrogram to perform SER. In these current trends, they have also discussed the growing scope of SERs like in the field of signal processing, pattern recognition, psychiatry. Also, for yielding higher success rates, the author speaks different machine learning methods are to be performed on various kinds of datasets having various kinds of tests.

Abhijit Mohanta et al. in [7] have analysed emotions such as angry, fear, happy, neutral from emotional speech signal where features such as loudness, detecting voiced region, excitation energy were used for analysis. They call these features as sub-segmental features. Using features such as, instantaneous fundamental frequency (F0) using Zero Frequency Filtering (ZFF), signal energy, formant frequencies, and dominant frequencies the analysis of these feelings has been done. The study analyses generation characteristics of four different emotion states and not the classification of emotional states portrayed by the actor. For locating instantaneous F0 and zero-crossing rate (ZCR), some Signal processing methods, ZFF and STE were used, respectively.

Edward Jones et al. in [8] considered Speech emotion recognition as an exciting ingredient of Human Computer Interaction (HCI). The main approach for SER must be feature extraction and feature classification. Linear and non-linear classifiers can be used for Feature classification. In linear classifiers, frequently used classifiers are Support Vector Machines (SVMs), Bayesian Networks (BN). Since, Speech signal is considered variating, thus, these types of classifiers work effectively for SER. Deep learning techniques possess more advantages for SER when compared to traditional methods. The deep learning techniques does not require manual feature extraction and tuning, also, it has capability to detect complex structure within.

Michael Neumann et al. in [9] presented their conclusions illustration gaining knowledge on unlabelled voice entity can be appropriate for Speech Emotion Recognition (SER). They have used t-distributed neighbour embeddings (t-SNE) to analyse visualizations of different representations. However, no divisible clusters are found in the 2D projections. These plots are excluded as of capacity they require. The autoencoder is trained on a large dataset. They have incorporated representations generated by autoencoders, which, in turn leads to steady developments in identification accuracy of SER model. The research also gives us a way to discover and experiment different alternatives of autoencoders and study procreative adversarial networks for representation learning.

Radim Burget et al. in [10] used German Corpus (Berlin Database of Emotional Speech) data which had more than 250 recordings. Each recording was distributed into 20 millisecond segments avoiding the overlapping. Then 3098 silent segments were eliminated to cut up the information into training, validation and testing sets. For removing the silent parts of audio Google WebRTC voice exercise detector was used in pre-processing. And then all the files are standardized so that they have zero mean and module variance. The input data given to Deep Neural Networks (DNN) were presented in batches with each of 21 iterations. Each batch contain the similar number of divisions and pattern is succeeded with pattern of different divisions like neutral, angry, sad and so on. DNN had no perception of the actual experience of what the actor is conveying, nor had any perception of the nature.

S.No	Title	Summary
1	Speech based Emotion Recognition using Machine Learning[1]	Was only limited to classify three emotions i.e., Angry, Sad and Happy
2	Speech Emotion Recognition Based on Deep Belief Network[2]	Emotions such as shame and surprise cannot be identified, affecting the overall rate of recognition and average rate of recognition
3	An Enhanced Human Speech Emotion Recognition Using Hybrid of PRNN and KNN[3]	Six universal feelings are classified like neutral, sadness, happiness, angry, fear and surprise over the given input.
4	Speech based human emotion recognition using MFCC[4]	Only one feature extraction is used i.e., MFCC. Due to which, only three emotions are confirmed irrespective of the gender
5	Research on Emergency Parking Instruction Recognition Based on Speech Recognition and Speech Emotion Recognition[5]	If the system recognises the feeling of the parking education expression, the efficiency is degraded due to the fact the voice stopping guidance database utilized in this paper are recorded totally extraordinary surroundings.
6	New Trends in Speech Emotion Recognition[6]	It obtained less accuracy. Changing dataset can prove as a solution to this problem.
7	Emotion recognition from speech signal[7]	It just analyses characteristics of various four emotion states, and does not include the classification of emotion states.
8	Speech Emotion Recognition Using Deep Learning Techniques: A Review[8]	The research work is done for different DNN techniques but no such implementation are done using this technique. It was just a theoretical concept.
9	Improving Speech Emotion Recognition with Unsupervised Representation Learning on Unlabeled Speech[9]	Representations became similar to factors. Henceforth, in 2D projections, separate clusters were not found which were bound to the space limitations
10	Speech emotion recognition with deep learning[10]	DNN had no understanding of the real sense of what the actor is try to saying, It. Neither it had any understanding of the speech, vibrations etc.

III. CONCLUSIONS

The Speech emotion recognition (SER) is a field that recognizes human emotions through the speech. A correct and precise database where the actors' voice is clear and noise free is ideal. An overview of SER methods is discussed for extracting audio features from speech sample, various classifier algorithms are used to recognise the emotion.

Various features are used to recognise emotions but feature extraction using MFCC seemed to have an important role in recognizing emotions through the speech. In this study, we saw various classifiers being used for classification. Whereas choosing the proper and best classifier is very important step in SER.

The accuracy of the Speech emotion recognition system is dependent upon the Database, the features extracted from the databases and a classification model (algorithm) used to classify the emotions.

REFERENCES

[1] Girija Deshmukh, Apurva Gaonkar, Gauri Golwalkar, Sukanya Kulkarni, "Speech based Emotion Recognition using Machine Learning", Institute Of Electrical And Electronics Engineers, Mar. 2019.

[2] Peng Shi, "Speech Emotion Recognition Based on Deep Belief Network", Institute Of Electrical And Electronics Engineers, March 2018.

[3] J. Umamaheswari, A. Akila, "An Enhanced Human Speech Emotion Recognition Using Hybrid of PRNN and KNN", Institute Of Electrical And Electronics Engineers, Feb 2019.

[4] Sri Raksha R. Gupta, M.S. Likitha, A. Upendra Raju and K. Hasitha "Speech Based Human Emotion Recognition Using MFCC", Institute Of Electrical And Electronics Engineers, March 2017.

[5] Tian Kexin, Huang Yongming, Zhang Guobao, Zhang Lin, "Research on Emergency Parking Instruction Recognition Based on Speech Recognition and Speech Emotion Recognition", Institute Of Electrical And Electronics Engineers, Nov. 2019.

[6] Ye Sim Ülgen Sonmez, Asaf Varol, "New Trends in Speech Emotion Recognition", Institute Of Electrical And Electronics Engineers, June 2019.

[7] Esther Ramdinmawii, Abhijit Mohanta, Vinay Kumar Mittal, "Emotion recognition from speech signal", Institute Of Electrical And Electronics Engineers, Nov. 2017.

[8] Ruhul Amin Khalil, Edward Jones, Mohammad Inayatullah Babar, Tariqullah Jan, Mohammad Haseeb Zafar, And Thamer Alhussain, "Speech Emotion Recognition Using Deep Learning Techniques: A Review", Institute Of Electrical And Electronics Engineers, Aug. 2019

[9] Michael Neumann, Ngoc Thang Vu, "Improving Speech Emotion Recognition with Unsupervised Representation Learning on Unlabeled Speech", Institute of Electrical and Electronics Engineers, May 2019.

[10] PavolHarár, RadimBurget, Malay Kishore Dutta, "Speech emotion recognition with deep learning", Institute Of Electrical And Electronics Engineers, Feb. 2017.