International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 04 Issue: 09 | Sept -2020 ISSN: 2582-3930

SPEECH EMOTION RECOGNITION

Dr. Praveen Shrivastava[1], Dr. Jitendra Sheetlani[2] , Ms.Puja Agrawal[3], Ms.Heena Aswani[4]

[1] Professor, Department of Computer Science, Sadhu Vaswani College, Bhopal
[2] Professor, Department of Computer Science, Sri Satya Sai University of Technical andMedical Sciences, Sehore
[3]Astt. Prof. , Department of Computer Science, Sadhu Vaswani College, Bhopal
[4]Astt. Prof. , Department of Computer Science, Sadhu Vaswani College, Bhopal

Abstract— The challenging module in CAS (computer-aided services) has recognized the emotion from the signals of speech. In SER (speech emotion recognition), several schemes have been used for extracting emotions from the signals, comprising various classifications & speech analysis methods. This paper presents an outline of methods & explores the existing models which have been used for emotion recognition based on speech. The database for the speech emotion recognition system comprises of emotional speech samples and the features extracted from these speech samples are the energy, pitch, linear prediction cepstrum coefficient (LPCC), Mel frequency cepstrum coefficient (MFCC). Inference about the performance and limitation of speech emotion recognition system based on the different classifiers are also discussed.

***Keywords*—**Computer-Aided Services, Speech Emotion recognition, Feature extraction, Automatic Speech Recognition, Natural Language Processing, MFCC.

I. INTRODUCTION

There are many ways of communication but the speech signal is one of the fastest and most natural methods of communication between human beings. Moreover the speech can be the fast and efficient method of interaction between human and machine also. The emotional detection is natural for humans but it is very difficult task for machine, thus the purpose of emotion recognition system is to use emotion related knowledge in such a way that human machine communication will be improved. Regardless of the remarkable enhancement made in understanding natural language & speech, still we are not capable of communicating with the machines naturally. Designing a method for understanding human emotions could be dominant for several human interactions of computer applications. Nevertheless, it could be very challenging to design such methods.

There are several modalities for expressing human emotions like body-posture, facial expression & voice. Hence, using manifold modalities might capture the expressed emotions accurately and result in optimal outcomes of recognition than uni-modal methods. Several contributions concentrated on utilizing modalities of audio-visual for recognition of emotion. This is because both of them were very significant features of the expression of emotion. Moreover, in conditions where only information of speech is available, one might use ASR

(Automatic speech recognition)[12] mechanism for converting signals of audio into text. Then, the multimodal method has applied for learning emotion from text & speech instantaneously. Hence, the text data were formed by the Automatic Speech Recognition (ASR) [12] mechanism that is trained generally from another

huge dataset for speech recognition. For effectively utilizing both text data & speech, one requires to devise a method for collectively learning features from distinct fields. Even though some of the researches integrated both trainings & features a multimodal method, some of the contributions concentrated on the temporal association among text & speech at an optimal level. In the speech recognition method from the end-end, the method employs a technique that has a word to be decoded to get its resulting frames of speech. By this contribution, the focus is on learning the alignment between text & speech.

II. SPEECH EMOTION RECOGNITION SYSTEM

Speech emotion recognition is nothing but the pattern recognitionsystem. This implies that the stages that are present in the pattern recognition system[7,13] are also present in the Speech emotion recognition system. The speech emotion recognition system contains five main modules emotional speech input, feature extraction, feature selection, classification, and recognized emotional output [5, 8].

The need to find out a set of the significant emotions to be classified by an automatic emotion recognizer is a main concern in speech emotion recognition system. A typical set of emotions contains 300 emotional states. Therefore to classify such a great number of emotions is very complicated. According to 'Palette theory', any emotion can be decomposed into primary emotions which are anger, disgust, fear, joy, sadness and surprise.The evaluation of the speech emotion recognition [5,8] system is based on the level of naturalness of the database which is used as an input to the speech emotion recognition system. If the inferior database is used as an input to the system then incorrect conclusion may be drawn. The database as an input to the speech emotion recognition system may contain the real world emotions or the acted ones. It is more practical to use database that is collected from the real life situations.

The conventional Speech Emotion Recognition methods usually include several classification methods like HMMs & GMMs. Here, GMMs are used to depict sound units of acoustic features. On other dimensions, HMMs are used to deal with the temporal changes in signals of speech. Here, the modeling procedure utilizing such conventional methods needs a huge dataset for attaining accuracy in recognizing emotion which would be consuming more time.

III. FEATURE EXTRACTION

Any emotion from the speaker's speech is represented by the large number of parameters which is contained in the speech and the changes in these parameters will result in corresponding change in emotions. Therefore an extraction of these speech features which represents emotions is an important factor in speech emotion recognition system[13]. The speech features can be divide into two main categories that is long term and short term features. Research on emotion of speech indicates that pitch, energy, duration, formant, Mel frequency cepstrum coefficient (MFCC) [10], and linear prediction cepstrum coefficient (LPCC) are the important features.

The prosodic features of voice like vitality & pitch were disengaged from every edge and known as highlights. Later, the worldwide highlights were defined as overall highlights of voice signal measurements that have eradicated from expression. Here, there has been an inconsistency on which worldwide highlights andneighborhoods are appropriately increasing for acknowledgment of voice signaling. The specialists have assured that the highlights of world-wide are effective in detecting excited feelings for example dread, euphoria, outrage, trouble and many more. They assure that the world-wide features disregard to cluster feelings that possess comparative excitement like happiness versus outrage. With the different emotional state, corresponding changes occurs in the speak rate, pitch, energy, and spectrum. Typically anger has a higher mean value and variance of pitch and mean value of energy. In the happy state there is an improvement in mean value, variation range and variance of pitch and mean value of energy. On the other hand the mean value, variation range and variance of pitch is decreases in sadness, also the energy is weak, speak rate is slow and decrease in spectrum of high frequency components. The feature of fear has a high mean value and variation range of pitch, improvement of spectrum in high frequency components.One of the main speech features which indicate emotion is energy and the study of energy is depends on short term energy and short term average amplitude.

In feature extraction [13] all of the basic speech feature extracted may not be helpful and essential for speech emotion recognition system. If all the extracted features gives as an input to the classifier this would not guarantee the best system performance which shows that there is a need to remove such a less useful features from the base features. Therefore there is a need of systematic feature selection to reduce these features. Forward selection method could be used to select the best feature subset. In the initial stage, forward selection initializes with the single best feature out of the whole feature set. The remaining features are further added which increases the classification accuracy.

IV. CLASSIFIER SELECTION

In the speech emotion recognition system after calculation of the features, the best features are provided to the classifier. A classifier recognizes the emotion in the speaker's speech utterance. Various types of classifier have been proposed for the task of speech emotion recognition. Gaussian Mixtures Model (GMM) [4], K-nearest neighbors (KNN), Hidden Markov Model (HMM)and Support Vector Machine (SVM), Artificial Neural Network (ANN) [9], etc. are the classifiers used in the speech emotion recognition system. Each classifier has some advantages and limitations over the others.

Gaussian Mixture Model (GMM) [4]is more suitable when global features are to be extracted for speech emotion recognition. All the training and testing equations are based on the supposition that all vectors are independent therefore GMM [4] cannot form temporal structure of the training data. Other classifier that is used for the emotion classification is an artificial neural network (ANN) [9], which is used due to its ability to find nonlinear boundaries separating the emotional states. Out of the many types, feed forward neural network is used most frequently in speech emotion recognition. According to the emotional state of the k utterances, the k-nearest neighbor classifier (K-NN) allocates an utterance to an emotional condition. The classifier can classify all the utterances in the design set properly, if 'k' equals to 1, however its performance on the test set will reduced. Hidden Markov Model (HMM) [11] is generally used for isolated word recognition and speech emotion recognition, the main reason is its physical relation with the speech signals production mechanism. In speech emotion recognition system, HMM has achieved great success for modeling temporal information in the speech spectrum.The support vector machine (SVM) [2,4]classifieris more capable of getting

optimum classification in the new feature space. SVMclassifier are generally used in the main applications like pattern recognition [7,13] and classification problems as well as in the speech emotion recognition system. SVM is having much better classification performance compared to other classifiers. The emotional states can be separated to huge margin by using SVM classifier. This margin is nothing but the width of the largest tube without any utterances, which can obtain around decision boundary.

V.OBSERVATION AND RESULTS

This section presents observations obtained from the contribution of this paper, which is the theoretical analysis of the features of speech signal towards emotion representation. The observations learned from the review carried on contemporary literature of emotion detection from speech signal include detection accuracy of the contemporary methods of emotion detection from speech signals to be very low, considerable false alarming in emotion detection from speech signals and absence of any kind of addressing of the crux of the volume and dimensionality of the training corpus.

Hence the significant objectives of the research are as follows:

- The robust machine learning strategies for speech signal recognition.
- Emotions centric features extraction and optimization.
- Handling the crux of volume and dimensionality of the training corpus.
- Building comprehensive classifiers to perform supervised learning and predictive analytics towards emotion detection from speech signals.

VI. CONCLUSION

Enhancement of the robustness [1] of emotion recognition system is still possible bycombining databases [13] and by fusion of classifiers. The effect of training multipleemotion detectors can be investigated by fusing these into a single detection system. The important issues in speech emotion recognition system are the signal processing unit in which appropriate features are extracted from available speech signal and another is a classifier which recognizes emotions from the speech signal. We aim also to use other feature selection methods because the quality of thefeature selection affects the emotion recognition rate.

REFERENCES

1. Lakomkin E, Zamani MA, Weber C, Magg S, Wermter S. On the Robustness of Speech Emotion Recognition for Human-Robot Interaction with Deep Neural Networks. In2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) 2018 Oct 1 (pp. 854-860). IEEE.

2. Zhang W, Zhao D, Chai Z, Yang LT, Liu X, Gong F, Yang S. Deep learning and SVM-based emotion recognition from Chinese speech for smart affective services. Software: Practice and Experience. 2017 Aug;47(8):1127-38.
3. Badshah AM, Ahmad J, Rahim N, Baik SW. Speech emotion recognition from spectrograms with deep convolutional neural network. In2017 international conference on platform technology and service (PlatCon) 2017 Feb 13 (pp. 1-5). IEEE.
4. Divya Sree GS, Chandrasekhar P,Venkateshulu B. SVM based speech emotion recognition compared with GMM-UBM and NN. IJESC. 2016;6.
5. Mao Q, Xue W, Rao Q, Zhang F, Zhan Y. Domain adaptation for speech emotion recognition by sharing priors between related source and target classes. In2016 IEEE international conference on acoustics, speech and signal processing (ICASSP) 2016 Mar 20 (pp. 2608-2612). IEEE.
6. Matilda S. Emotion recognition: A survey. International Journal of Advanced Computer Research. 2015;3(1):14-19.
7. Albornoz EM, Sánchez-Gutiérrez M, Martinez-Licona F, Rufiner HL, Goddard J. Spoken emotion recognition using deep learning. InIberoamerican Congress on Pattern Recognition 2014 Nov 2 (pp. 104-111). Springer, Cham.
8. Huang Z, Dong M, Mao Q, Zhan Y. Speech emotion recognition using CNN. InProceedings of the 22nd ACM international conference on Multimedia 2014 Nov 3 (pp. 801-804). ACM.
9. Yu D, Seltzer ML, Li J, Huang JT, Seide F. Feature learning in deep neural networks-studies on speech recognition tasks. arXiv preprint arXiv:1301.3605. 2013 Jan 16.
10. Milton A, Sharmy Roy S, Tamil Selvi S. SVM scheme for speech emotion recognition using MFCC feature.International Journal of Computer Applications. 2013;69.
11. Ingale AB, Chaudhari D. Speech emotion recognition using hidden Markov model and support vector machine. International Journal of Advanced Engineering Research and Studies. 2012:316-318.
12. Peipei S, Zhou C, Xiong C.Automatic speech emotion recognition using support vector machine. IEEE. 2011;2:621-625.
13. El Ayadi M, Kamel MS, Karray F. Survey on speech emotion recognition: Features, classification schemes, and databases. Pattern Recognition. 2011 Mar 1;44(3):572-87.
14. Huang KC, Kuo YH. A novel objective function to optimize neural networks for emotion recognition from speech patterns. In2010 Second World Congress on Nature and Biologically Inspired Computing (NaBIC) 2010 Dec 15 (pp. 413-417). IEEE.