iJRASET
International Journal For Research in
Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 9 **Issue:** I **Month of publication:** January 2021

DOI: https://doi.org/10.22214/ijraset.2021.32740

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 9 Issue I Jan 2021- Available at www.ijraset.com

Speech Emotion Recognition

Ramya. T[1], Erica Davey[2], Kavya. M[3], Uzma Taj[4], Mrs. Supriya[5]
[1, 2, 3,4]Student, [5]Guide, Department of CSE, Sir MVIT

Abstract: Recognizing emotions is automatically and subconsciously performed by humans. It is a vital process for human-to human communication, and thus, to achieve better human machine interaction, emotions need to be considered. Emotional speech recognition importance is growing in several domains. Researchers have raised the impact of emotion in multidisciplinary applications. Predicting human emotions is catching the attention of many research areas, which demand accurate predictions in uncontrolled scenarios. Psychologists have widely studied the influence of emotional factors, on decision-making. As example, pilot's decision in a flight context may jeopardize several humans life . The importance of recognizing emotion for needs of the real world use is becoming unavoidable. In real world applications, speech signals are often corrupted by acoustic background noise. In these applications, speech enhancement is a necessary module for the emotion recognition system. Despite recent advances in the field of automatic speech emotion recognition, recognize emotions from vocal channel in noisy environment remains an open research problem. We proposed a system that will do Speech detection and Continuous recognition of emotions from speech.

I. INTRODUCTION

Language is the most natural way of communication. The emotional signal in the voice signal is one of the important information expressions, and is the necessary part of the information in the human perception. In a system comprising of human-machine interaction, the emotion recognition of speech has always been a wide area of research since the machines can never analyze the emotion of a speaker on its own. The mode of speech is the quickest and the most characteristic strategy for correspondence between people. This reality has persuaded researchers to consider speech as a quick and effective strategy for communication amid humans as well as machines. Conversely, this necessitates the machine ought to have adequate insight to perceive human voices. .Since the late fifties, there has been enormous research over recognition of speech modes, which alludes to the development in changing over the human discourse into a grouping of words. However, in spite of an extraordinary advancement made in recognition of speech, we are still a long way from having a characteristic association among man and machine because the machine does not comprehend the emotional condition of the speaker. Affect recognition is an important component towards the better interaction between human and machines. Applications of emotion recognition in speech can be found in several areas such as human computer interaction and call centres. Recognizing emotions is automatically and subconsciously performed by humans. It is a vital process for human-to-human communication, and thus, to achieve better human machine interaction, emotions need to be considered. There are three major approaches for quantifying emotions, namely, categorical, continuous and appraisal- based.

Deep Neural Networks (DNNs) have emerged the recent years and had groundbreaking improvements in different areas of machine learning including the continuous affect. The authors share joint first authorship recognition domain. Numerous new DNNs architectures have been proposed recently towards that direction such as Convolutional Neural Networks (CNNs), and Long-Short Term Memory (LSTM) networks. Speech is a complex signal consisting of various information, such as information about the message to be communicated, speaker, language, region, emotions etc. Speech Processing is one of the important branches of digital signal processing and finds applications in Human computer interfaces, Telecommunication, Assistive technologies, Audio mining, Security and so on. Speech emotion recognition is important to have a natural interaction between human being and machine. In speech emotion recognition, emotional state of a speaker is extracted from his or her speech. The acoustic characteristic of the speech signal is Feature. Feature extraction is the process that extracts a small amount of data from the speech signal that can later be used to represent each speaker.

We proposed a system that will do Speech detection and Continuous recognition of emotions from speech.

II. LITERATURE SURVEY

J. Umamaheswari and A. Akila. [1] In this proposed technique, an enhanced speech emotion recognition is carried out over six basic emotions of angry, happy, sad, neutral, surprise and fear. Here, as an advanced research methodology, the pre-processing was carried out using PRNN and KNN algorithms while the feature extraction was made using a cascaded structure comprising of MFCC and GLCM.

©IJRASET: All Rights are Reserved

In this research study, an enhanced human speech emotion recognition system using a hybrid of PRNN and KNN algorithms is designed. The six basic emotions like neutral, anger, happiness, sadness, surprise and fear over the speech emotions are classified and studied for their accuracy with other previously developed systems. The database for this study is taken as the emotional speech samples of numbers. A cascaded system of Mel Frequency Cepstral Coefficient (MFCC) and Gray Level Co-occurrence Matrix (GLCM) was used for feature extraction process along with a Wiener filter for filtering the noise in speech. Also, a hybrid of Pattern Recognition Neural Network (PRNN) and KNearest Neighbor (KNN) is used for prediction accuracy of outcomes. The outcomes are compared with previously developed recognition systems and better efficiency is observed. The obtained results were compared for their accuracy, precision rate and f-Measure with standard algorithms like GMM and HMM and were recognized as a better output than the standard algorithms.

H. Zheng and Y. Yang. [2] In order to improve the characterization ability of speech signal and recognition accuracy of speech emotion recognition, a speech emotion recognition model based on improved Deep Belief Network (DBN) is proposed. The method is to replace the traditional DBN activation function with a Rectified Linear Unit (Relu). And the reconstruction error is used to determine the depth of the DBN network. The short time energy, short-time zero crossing rate, the fundamental frequency, formants and 24 dimensional MFCC parameters of emotional speech signal are extracted as the basic features. Using these basic features as input to the DBN, automatic recognition of the 6 emotions, anger, fear, joy, calmness, sadness and surprise can be achieved. Compared with the traditional DBN model and the BP model, a better recognition result is achieved by using the improved DBN discussed in this paper, and the recognition rate can reach 84.94%. This paper introduces a speech emotion recognition system, and extract the short-term energy, short-time zero crossing rate, gene frequency, first resonance and MFCC. The results show that the eigenvalues selected in this paper are reasonable, and the method of extracting is effective and feasible. The Relu function replaces the sigmoid function of the traditional DBN, and the reconstruction error of the RBM training is improved, thereby improving the recognition efficiency. Reconstruction error is used to determine the number of hidden layers in DBN model, which makes the model more stable.

Michael Neumann, Ngoc Thang Vu. [3] In this paper, they have shown that incorporating representations generated by an autoencoder that was trained on a large dataset, leads to consistent improvements in recognition accuracy of the presented SER model. Further they presented t-SNE visualizations that reveal the discriminative strength of those representations with regard to low and high arousal. Future work includes experimentation with different variants of autoencoders and investigation in generative adversarial networks for representation learning. In this paper they present findings on how representation learning on large unlabeled speech corpora can be beneficially utilized for speech emotion recognition (SER). Prior work on representation learning for SER mostly focused on the relatively small emotional speech datasets without making use of additional unlabeled speech data. They show that integrating representations learnt by an unsupervised autoencoder into a CNN-based emotion classifier improves the recognition accuracy. To gain insights about what those models learn, we analyze visualizations of the different representations using t- distributed neighbor embeddings (t-SNE). We evaluate our approach on IEMOCAP and MSP- IMPROV by means of within and cross-corpus testing.

Valery A. Petrushin. [4] The paper describes two experimental studies on vocal emotion expression and recognition. The first study deals with a corpus of 700 short utterances expressing five emotions: happiness, anger, sadness, fear, and normal (unemotional) state, which were portrayed by thirty non-professional actors. After evaluation a part of this corpus was used for extracting features and training backpropagation neural network models. Some statistics of the pitch, the first and second formants, energy and the speaking rate were selected as relevant features using feature selection techniques. Several neural network recognizers and ensembles of recognizers were created. The recognizers have demonstrated the following accuracy: normal state - 60-75%, happiness – 60- 70%, anger – 70-80%, sadness – 70-85%, and fear – 35-55%. The total average accuracy is about 70%. The second study uses a corpus of 56 telephone messages of varying length (from 15 to 90 seconds) expressing mostly normal and angry emotions that were recorded by eighteen non-professional actors. These utterances were used for creating recognizers using the methodology developed in the first study. The recognizers are able to distinguish between two states: “agitation” which includes anger, happiness and fear, and “calm” which includes normal state and sadness with the average accuracy 77%. An ensemble of such recognizers was used as a part of a decision support system for prioritizing voice messages and assigning a proper agent to response the message. In this paper they explored how well people and computers recognize emotions in speech. Several conclusions can be drawn from the above results. First, decoding of emotions in speech is complex process that is influenced by cultural, social, and intellectual characteristics of subjects. People are not perfect in decoding even such manifest emotions as anger and happiness. Second, anger is the most recognizable and easier to portray emotion. It is also the most important emotion for business.

©IJRASET: All Rights are Reserved

Farah Chenchah and Zied Lachiri. [5] In this paper, they proposed an emotion recognition approach having as objective to recognize emotion on real life condition with four types of noise. We added to the classical emotion recognition system a speech enhancement step using spectral substraction, wiener filter and MMSE. We noticed that spectral substraction and MMSE enhance significantly the recognition rate in airport and babble noise, and that the proposed enhancement methods are not efficient to reduce car noise. Three speech enhancement algorithms are introduced for improved emotion classification; spectral subtraction, wiener filter and MMSE. Experiments were prepared with MFCC as feature vectors and HMM as classifier. Experiments are evaluated on real condition speech signal (IEMOCAP database) with real world noise using various SNR level. Results after denoising were compared to those before denoising and those without noise to measure the system performance. The experimental results show that the speech enhancement algorithms improve the performance of our emotion recognition system under various SNRs.

Tian Kexin, Huang Yongming, Zhang Guobao and Zhang Lin. [6] In this paper, dynamic time warping (DWT) algorithm is used to match the parking command distance. Firstly, the feature vector sequences of each parking instruction are matched by algorithm, the feature vector sequence with the smallest sum of matching distances is selected as a reference template, and the threshold of the matching distance is determined experimentally. Experiments show that the method is simple and the recognition rate is high. This paper proposes an emergency parking instruction recognition system based on speech recognition and speech emotion recognition. The system first extracts the feature vector of the speech signal, and then uses the support vector machine (SVM) to recognize the emotion of the speech. When the emotion is abnormal, the dynamic time warping (DWT) algorithm is used to match the parking instruction template. The test results show that the system can realize the speech recognition of parking instructions very well.

Zhiyan HAN and Jian WANG. [7] This paper used Gaussian Kernel Nonlinear Proximal Support Vector Machine to recognize the emotions, and improved the emotion recognition accuracy. But the way of human beings expressing emotions is diverse, it has the expression complexity and culture relative property. There are many limitations for only using speech to recognize emotion. So they can combine facial expression signal to recognize emotion. Of course, the characteristic parameters have great influence on the recognition results .

Elif Bozkurt, Engin Erzin and A. Tanju Erdem. [8] They propose the use of the line spectral frequency (LSF) features for emotion recognition from speech, which have not been been previously employed for emotion recognition to the best of our knowledge. Spectral features such as mel-scaled cepstral coefficients have already been successfully used for the parameterization of speech signals for emotion recognition. The LSF features also offer a spectral representation for speech, moreover they carry intrinsic information on the formant structure as well, which are related to the emotional state of the speaker . We use the Gaussian mixture model (GMM) classifier architecture, that captures the static color of the spectral features. Experimental studies performed over the Berlin Emotional Speech Database and the FAU Aibo Emotion Corpus demonstrate that decision fusion configurations with LSF features bring a consistent improvement over the MFCC based emotion classification rates. In this paper, they investigate the contribution of the line spectral frequency (LSF) features to the speech driven emotion recognition task. The LSF features are known to be closely related to the formant frequencies, however they have not been previously employed for emotion recognition to the best of our knowledge. We demonstrate through experimental results on two different emotional speech databases that the LSF features are indeed beneficial and bring about consistent recall rate improvements for emotion recognition from speech. In particular, the decision fusion of the LSF features with the MFCC features results in improved classification rates over the state-of-the-art MFCC-only decision for both of the databases. Olivier Lahaie, Roch Lefebvre and Philippe Gournay. [9] This paper presented a study of the influence of bandwidth limitation on the perception of speech emotions by human listeners. Several standard telephony bandwidths (fullband, superwideband, wideband, narrowband MSIN and narrowband IRS) were considered. The results showed that, in some cases, recognition accuracy decreases with audio bandwidth. More importantly, the number of listenings before subjects made a decision increased as the bandwidth decreased, which indicates that bandwidth limitation makes the recognition task harder

III. CONCLUSION

Recognizing emotions is automatically and subconsciously performed by humans. It is a vital process for human-to human communication, and thus, to achieve better human machine interaction, emotions need to be considered. The Automated Speech Emotion Recognition is a tough process because of the gap among acoustic characteristics and human emotions, which depends strongly on the discriminative acoustic characteristics extracted for a provided recognition task. Different persons have different emotions and altogether a different way to express it. Speech emotion do have different energies, pitch variations are emphasized if considering different subjects. Therefore, the speech emotion detection is a demanding task in computing vision.

©IJRASET: All Rights are Reserved

Here, the speech emotion recognition is based on the Convolutional Neural Network (CNN) algorithm which uses different modules for the emotion recognition and the classifiers are used to differentiate emotions such as happiness, surprise, anger, neutral state, sadness, etc. Emotional speech recognition importance is growing in several domains. So a solution to this we proposed a system that will do Speech detection and Continuous recognition of emotions from speech.

REFERENCES

[1] M. Neumann and N. T. Vu, "Improving Speech Emotion Recognition with Unsupervised Representation Learning on Unlabeled Speech," ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, United Kingdom, 2019, pp. 7390-7394, doi: 10.1109/ICASSP.2019.8682541.

[2] T. Kexin, H. Yongming, Z. Guobao and Z. Lin, "Research on Emergency Parking Instruction Recognition Based on Speech Recognition and Speech Emotion Recognition," 2019 Chinese Automation Congress (CAC), Hangzhou, China, 2019, pp. 2933-2937, doi: 10.1109/CAC48633.2019.8997077.

[3] J. Umamaheswari and A. Akila, "An Enhanced Human Speech Emotion Recognition Using Hybrid of PRNN and KNN," 2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon), Faridabad, India, 2019, pp. 177-183, doi: 10.1109/COMITCon.2019.8862221.

[4] F. Chenchah and Z. Lachiri, "Speech emotion recognition in noisy environment," 2016 2nd International Conference on Advanced Technologies for Signal and Image Processing (ATSIP), Monastir, 2016, pp. 788-792, doi: 10.1109/ATSIP.2016.7523189.

[5] Z. Han and J. Wang, "Speech emotion recognition based on Gaussian kernel nonlinear proximal support vector machine," 2017 Chinese Automation Congress (CAC), Jinan, 2017, pp. 2513-2516, doi: 10.1109/CAC.2017.8243198.

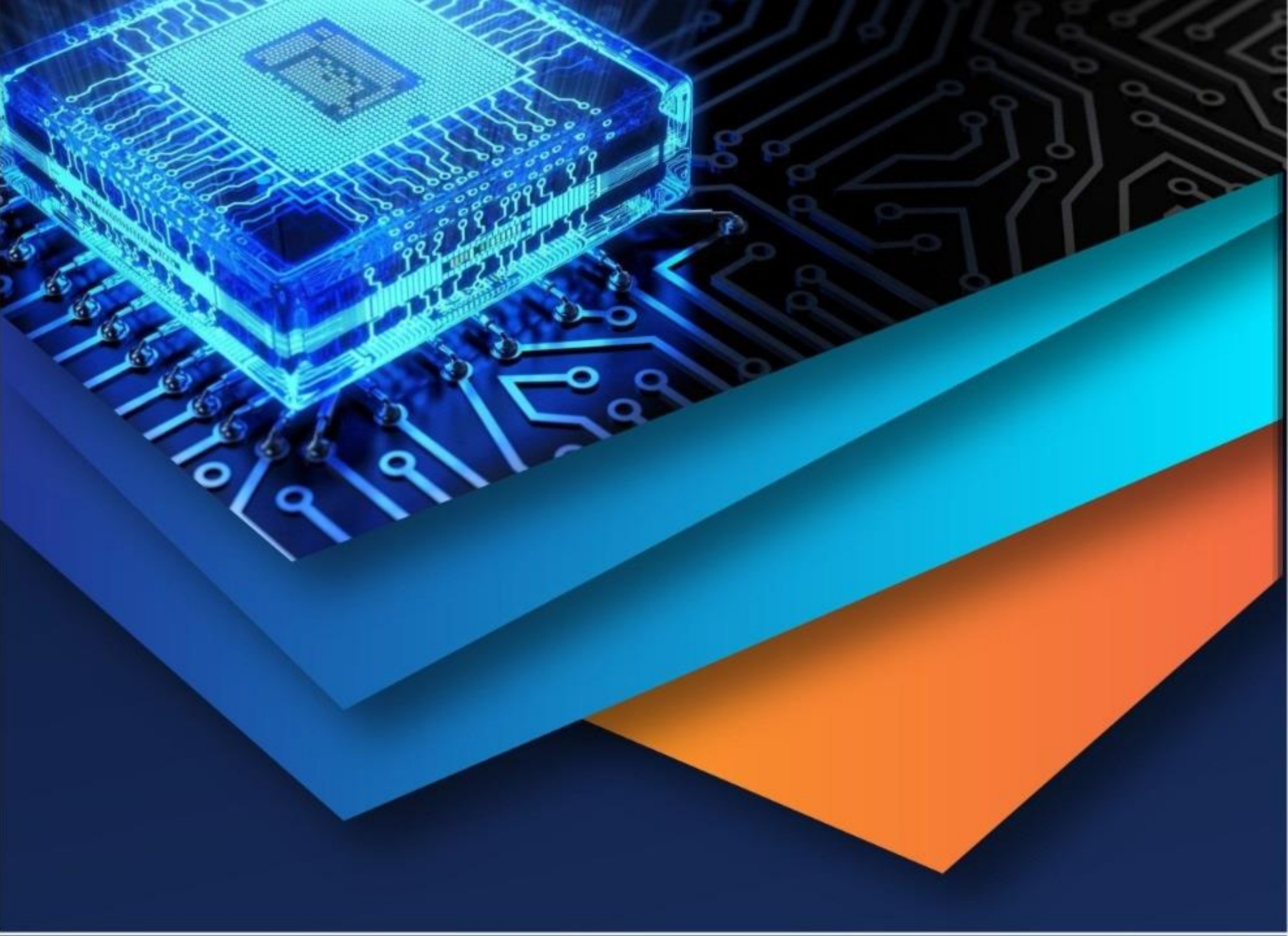
[6] E. Bozkurt, E. Erzin, C. E. Erdem and A. T. Erdem, "Use of Line Spectral Frequencies for Emotion Recognition from Speech," 2010 20th International Conference on Pattern Recognition, Istanbul, 2010, pp. 3708-3711, doi: 10.1109/ICPR.2010.903.

[7] H. Zheng and Y. Yang, "An Improved Speech Emotion Recognition Algorithm Based on Deep Belief Network," 2019 IEEE International Conference on Power, Intelligent Computing and Systems (ICPICS), Shenyang, China, 2019, pp. 493-497, doi: 10.1109/ICPICS47731.2019.8942482.

[8] R. Lotfidereshgi and P. Gournay, "Biologically inspired speech emotion recognition," 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), New Orleans, LA, 2017, pp. 5135-5139, doi: 10.1109/ICASSP.2017.7953135.

[9] O. Lahaie, R. Lefebvre and P. Gournay, "Influence of audio bandwidth on speech emotion recognition by human subjects," 2017 IEEE Global Conference on Signal and Information Processing (GlobalSIP), Montreal, QC, 2017, pp. 61-65, doi: 10.1109/GlobalSIP.2017.8308604.

©IJRASET: All Rights are Reserved



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)