EMOTION RECOGNITION USING SPEECH PROCESSING

DR K VENKATA SUBBAIAH [1],

Ch. Lekhya Harshika [2], T. Veera Ashrith Mehar [3], Ch. Lakshmi [4],

G. Yaswin Sai Goud [5]

#1Professor in Department of CSE-AI, PBR Visvodaya Institute of Technology and Science, Kavali.

#2#3#4#5B.Tech with Specialization of Computer Science and Engineering-Artificial Intelligence in PBR Visvodaya Institute of Technology and Science, Kavali.

ABSTRACT_ Emotion recognition from speech signals has received a lot of interest in the field of human-machine interaction applications, and it's been a key research area for many years. Emotions have enormous influence on human mental processes, serving as conduits for the communication of personal perspectives and internal moods to others. Speech Emotion Recognition (SER) is the act of determining speakers' emotional states from their speech signals, representing a complex endeavor aimed at comprehending the intricate intricacies of human expression.

This study aimed to dive into the complex fabric of human emotions using a comprehensive mix of signal processing techniques and machine learning algorithms. The project aims to capture the wide spectrum of emotional cues encoded in speech signals by employing a variety of feature extraction approaches such as Mel-frequency cepstral coefficients (MFCC), chromogram, Mel-scaled spectrogram, spectral contrast, and tonal centroid features. By combining these qualities, the project hopes to provide a comprehensive representation of emotional states, allowing for the classification and synthesis of universal emotions like neutrality, anger, happiness, and sadness.

.At the heart of this study is the use of Deep Neural Networks (DNNs) as powerful tools for emotion classification. The study aims to untangle the subtle interplay between acoustic parameters and emotional states contained in voice signals by leveraging the expressive capacity of deep neural networks. The DNN model aims to develop a strong capacity for discerning and categorizing a wide range of emotional expressions through rigorous training and validation processes, paving the way for improved human-machine interaction paradigms based on empathetic understanding and nuanced communication.

1.INTRODUCTION

There are numerous modes of communication, but the speech signal is one of the quickest and most natural forms of communication between humans. As a result, speech can be a quick and effective way for humans and machines to communicate. Humans have the innate ability to employ all of their senses to get maximum awareness of the message they hear. People perceive their communication partner's emotional condition using all of their senses. Humans are naturally good at detecting emotions, while machines struggle with it. As a result, the goal of an emotion detection system is to improve human-machine communication by utilizing emotion-related knowledge.

.

In this system, feature extraction quality has a direct impact on voice emotion identification accuracy. During the feature extraction process, the entire emotion sentence was usually used as a unit for feature extraction, and the extraction contents were four aspects of emotion speech: time construction, amplitude construction, fundamental frequency construction, and formant construction. Then contrast emotion speech with no emotion sentence from these four features, getting the law of emotional signal distribution, and classifying emotion speech according to the law.

.

Deep neural networks (DNNs) have achieved extraordinary success in voice recognition and picture recognition; nevertheless, no research on deep neural networks has been conducted on speech emotion processing. We discovered that the DNN outperforms other methods for processing speech emotions. As a result, this research offered a method for automatically extracting emotional aspects from audio using Python's librosa library. We utilized DNN to train a 5-layer deep network to extract speech emotion characteristics. It employs the speech emotion data of multiple consecutive frames to create a high latitude characteristic and a softmax classifier layer to categorize emotional speech. The speech emotion recognition test accuracy reached 73.38%, which is a high value in comparison to the others. To identify emotions, traditional machine learning algorithms include k-nearest neighbors (KNN), Hidden Markov Model (HMM), Support Vector Machine (SVM), Artificial Neural Network (ANN), Gaussian Mixtures Model (GMM), and so on.

The signal processing unit extracts appropriate features from available speech

signals, and another is a classifier that classifies emotions from the spoken signal. The average accuracy of most classifiers in a speaker-independent system is lower than in a speaker-dependent system. Automatic emotion recognition from human speech is becoming more popular since it leads to better interactions between humans and machines.

2.LITERATURE SURVEY

[1] Szegedy, Christian & Toshev, Alexander & Erhan, Dumitru. (2013). Deep Neural Networks for Object Detection. 1-9.

Title: Deep Neural Networks for Object Detection

Deep Neural Networks (DNNs) have recently shown outstanding performance on image classification tasks. In this paper we go one step further and address the problem of object detection using DNNs, that is not only classifying but also precisely localizing objects of various classes.

We present a simple and yet powerful formulation of object detection as a regression problem to object bounding box masks. We define a multi-scale inference procedure which is able to produce high-resolution object detections at a low cost by a few network applications. State-of-the-art performance of the approach is shown on Pascal VOC.

Summary: This journal discusses about the Deep Neural Networks theory and object detection using DNN.

[2] Benk, Sal & Elmir, Youssef & Dennai, Abdeslem. (2019). A Study on Automatic Speech Recognition. 10. 77-85. 10.6025/jitr/2019/10/3/77-85.

Title: A Study on Automatic Speech Recognition

Speech is an easy and usable technique of communication between humans, but nowadays humans are not limited to connecting to each other but even to the different machines in our lives. The most important is the computer. So, this communication technique can be used between computers and humans. This interaction is done through interfaces, this area called Human Computer Interaction (HCI). This paper gives an overview of the main definitions of Automatic Speech Recognition (ASR) which is an important domain of artificial intelligence and which should be taken into account during any related research (Type of speech, vocabulary size... etc.).

It also gives a summary of important research relevant to speech processing in the few last years, with a general idea of our proposal that could be

considered as a contribution in this area of research and by giving a conclusion referring to certain enhancements that could be in the future works.

Summary: This article helps us in understanding and using the speech recognition by machines which improves Human Computer Interactions and is also useful in our project.

[3] Ashish B. Ingale & D. S. Chaudhari (2012). Speech Emotion Recognition. International Journal of Soft Computing and Engineering (IJSCE) ISSN: 2231-2307, Volume-2 Issue-1, March 2012

Title: Speech Emotion Recognition from International Journal of Soft Computing and Engineering

In human machine interface application, emotion recognition from the speech signal has been research topic since many years. To identify the emotions from the speech signal, many systems have been developed. In this paper speech emotion recognition based on the previous technologies which uses different classifiers for the emotion recognition is reviewed. The classifiers are used to differentiate emotions such as anger, happiness, sadness, surprise, neutral state, etc.

The database for the speech emotion recognition system is the 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). The classification performance is based on extracted features. Inference about the performance and limitation of speech emotion recognition system based on the different classifiers are also discussed.

Summary: In this paper, we learn the importance and the need of a different features in any audio or speech including mfcc, mel and other features which are used in our application for the purpose of predicting the emotions based on audio

3.PROPOSED SYSTEM

We propose a modified voice emotion recognition approach based on deep neural networks for training. To extract information from an audio file, the approach employs Mel-frequency cepstral coefficients (MFCC), a chromogram, a Mel scaled spectrogram, as well as spectral contrast and Tonal Centroid characteristics. The features are utilized to train a DNN model with five layers. The dataset used in this study is the Ryerson

Audio-Visual Database of Emotional Speech and Song (RAVDESS). We simply chose the speaking segment, which includes 24 performers (gender balanced) and 1440 audio files. The algorithm classifies speech sounds into eight different emotions: neutral, calm, happy, sad, angry, afraid, disgusted, and startled.

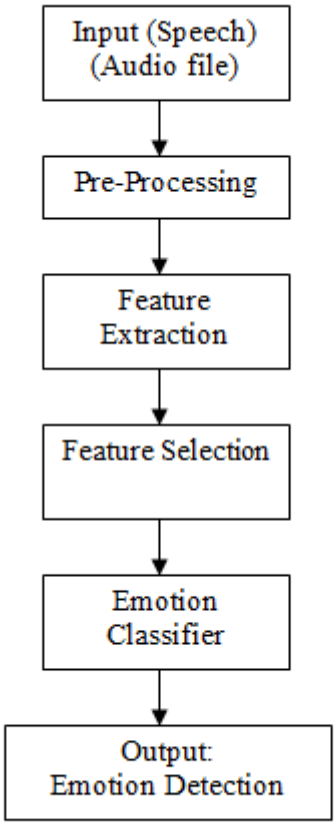


Fig 1:Architecture

3.1 IMPLEMENTATION

Data Collection Module:

This module involves gathering the audio data from the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS). Specifically, the speaking segments from 24 performers (gender balanced) and 1440 audio files are selected for analysis.

Preprocessing Module:

Preprocessing steps include segmenting the audio files, applying noise reduction techniques if necessary, and standardizing the audio format. This ensures consistency and prepares the data for feature extraction.

Feature Extraction Module:

This module extracts various acoustic features from the preprocessed audio files. Features include Mel-frequency cepstral coefficients (MFCC), chromograms, Mel-scaled spectrograms, spectral contrast, and Tonal Centroid characteristics. These features capture important information related to pitch, timbre, and spectral characteristics of the audio.

Model Training Module:

The extracted features are used to train a Deep Neural Network (DNN) model with five layers. The DNN architecture is designed to learn the complex patterns and relationships between the input features and the corresponding emotional labels. Training involves optimizing the model parameters using techniques like backpropagation and gradient descent.

Model Evaluation Module:

After training, the performance of the DNN model is evaluated using metrics such as accuracy, precision, recall, and F1-score. Cross-validation techniques may be employed to ensure the generalization ability of the model across different datasets.

User Interface Module:

A user-friendly interface is developed to allow users to interact with the emotion recognition system. This interface may include features for uploading audio files, displaying recognized emotions, and providing feedback.

Integration Module:

Once individual modules are developed and tested, they are integrated into a cohesive system. This module ensures seamless communication between different modules and the overall functionality of the emotion recognition system.

Deployment Module:

The final step involves deploying the developed system for real-world applications. This may include packaging the system for distribution, optimizing performance for different environments, and providing documentation for users.

4.RESULTS AND DISCUSSION

Fig 2:Registration Form:

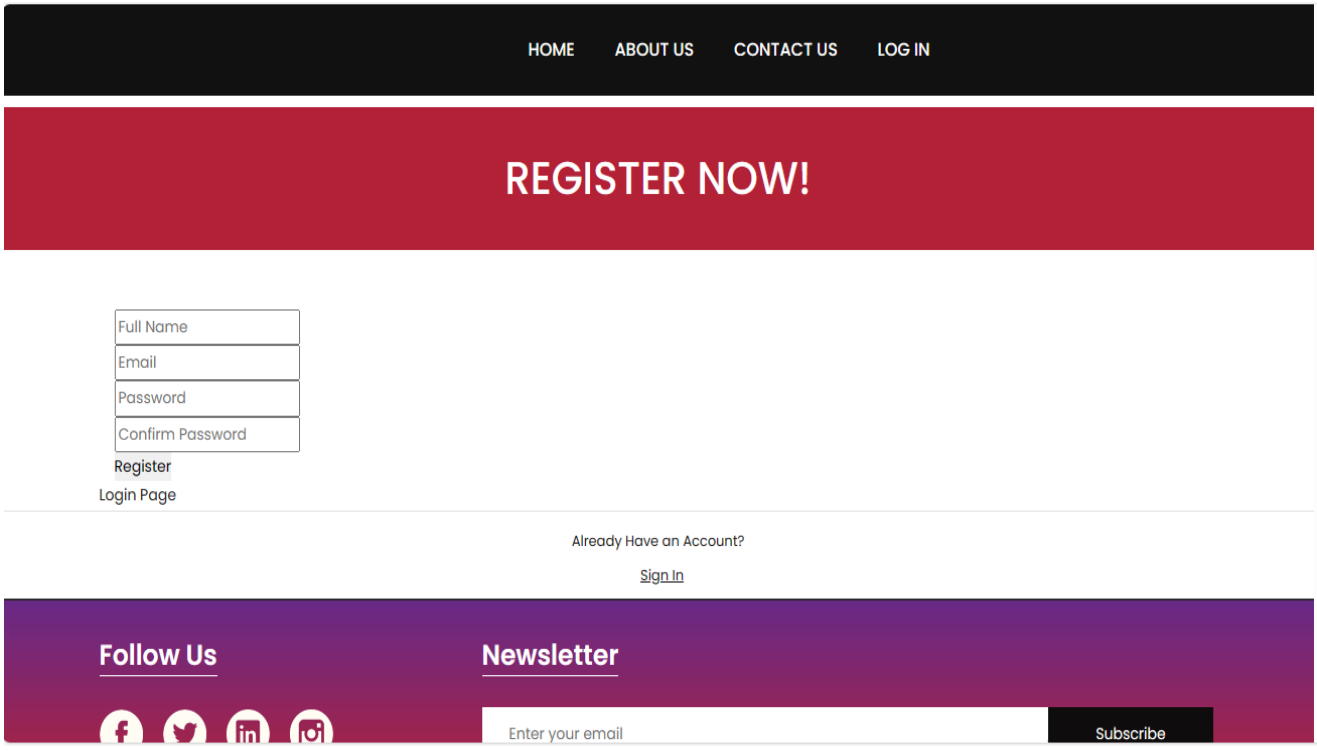


Fig 3:Login Form:

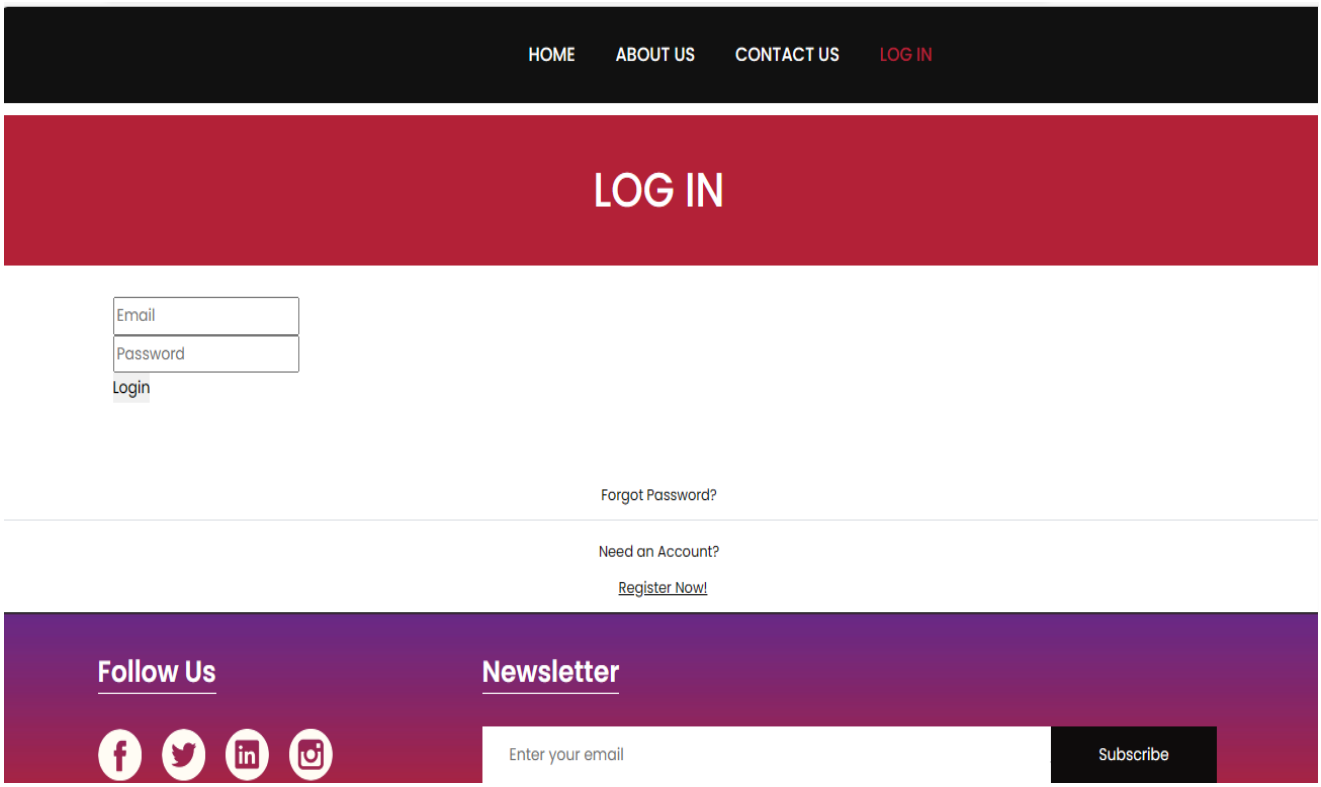


Fig 4:Input:

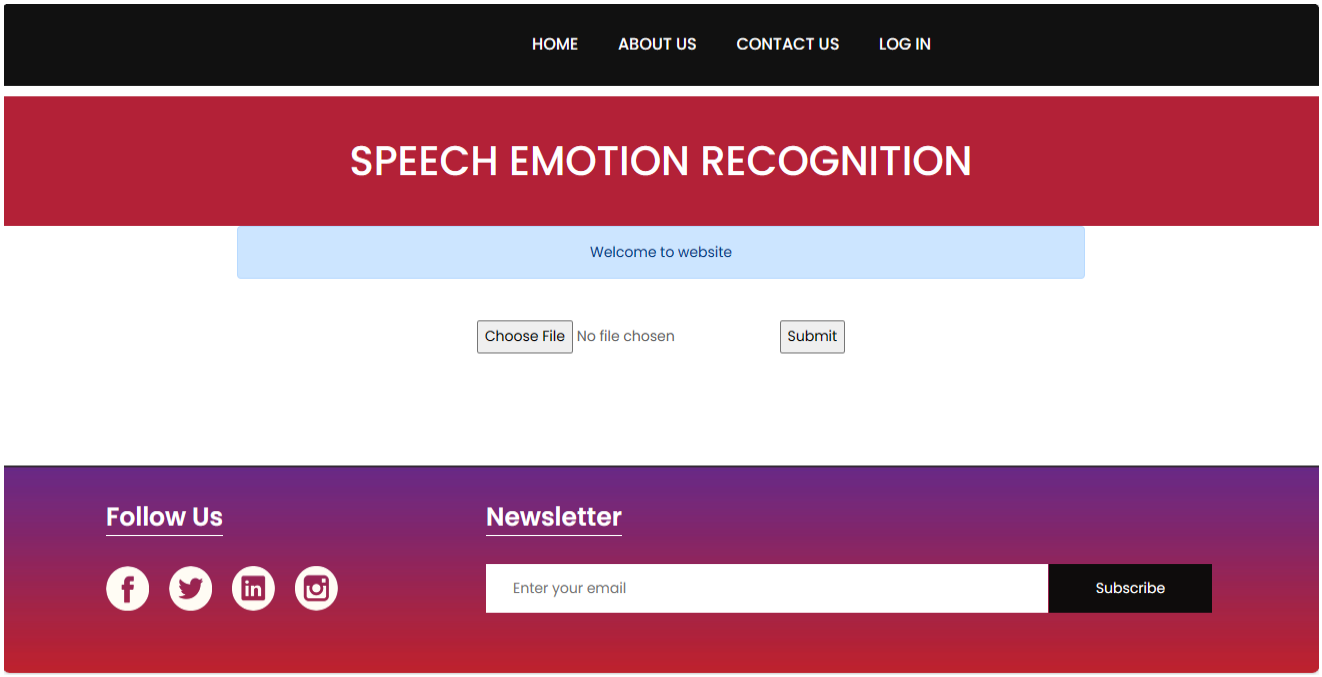
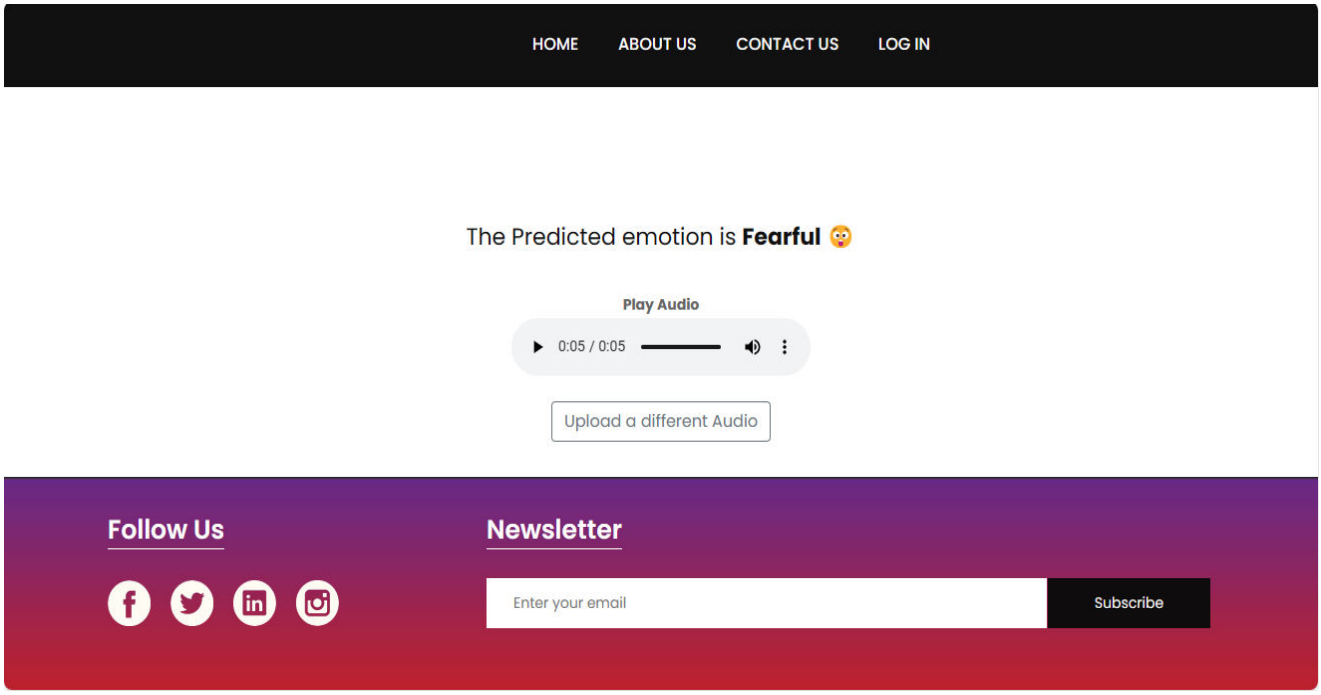


Fig 5:Output:



5.CONCLUSION

The proposed scheme presented an approach to recognize the emotion from the human speech. This approach has been implemented by the using the neural networks. We have successfully developed a deep learning model using the deep neural network architecture to predict the emotions of the speaker in an audio. We have famed our project in a web - based application using the Flask architecture. The UI also includes user registration system. We were able to get a test accuracy of 73.4% using the trained model.

Please note that emotion prediction is subjective and the emotions rated by a person for the same audio can differ from person to person. This is also the reason why the algorithm which is trained on human rated emotions can generate erratic results sometimes. The model was trained of RAVDESS dataset, so the accent of the speaker can also lead to erratic results as the model is only trained on North American accent database.

FUTURE SCOPE:

The ability to record a voice live and to predict the emotions in real time as the speaker is speaking. It also requires the knowledge of signal processing as the voice needs to be cleaned of all the unwanted noises in them before prediction.

REFERENCES

1. Szegedy, Christian & Toshev, Alexander & Erhan, Dumitru. (2013). Deep Neural Networks for Object Detection. 1-9.

2. Benk, Sal & Elmir, Youssef & Dennai, Abdeslem. (2019). A Study on Automatic Speech Recognition. 10. 77-85. 10.6025/jitr/2019/10/3/77-85.
3. Ashish B. Ingale & D. S. Chaudhari (2012). Speech Emotion Recognition. International Journal of Soft Computing and Engineering (IJSCE) ISSN: 2231-2307, Volume-2 Issue-1, March 2012
4. Chenchen Huang, Wei Gong, Wenlong Fu, Dongyu Feng, "A Research of Speech Emotion Recognition Based on Deep Belief Network and SVM", Mathematical problems in Engineering, vol. 2014 Article ID 749604, 7 pages, 2014. https://doi.org/10.1155/2014/749604
5. Livingstone SR, Russo FA (2018) The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. PLoS ONE 13(5): e0196391.
6. M. E. Ayadi, M. S. Kamel, F. Karray, "Survey on Speech Emotion Recognition: Features, Classification Schemes, and Databases", Pattern Recognition 44, PP.572-587, 2011.
7. I. Chiriacescu, "Automatic Emotion Analysis Based On Speech", M.Sc. THESIS Delft University of Technology, 2009.
8. T. Vogt, E. Andre and J. Wagner, "Automatic Recognition of Emotions from Speech: A review of the literature and recommendations for practical realization", LNCS 4868, PP.75-91, 2008.
9. S. Emerich, E. Lupu, A. Apatean, "Emotions Recognitions by Speech and Facial Expressions Analysis", 17th European Signal Processing Conference, 2009.
10. P.Shen, Z. Changjun, X. Chen, "Automatic Speech Emotion Recognition Using Support Vector Machine", International Conference On Electronic And Mechanical Engineering And Information Technology, 2011.

Author's Profiles

DR K VENKATA SUBBAIAH received his Ph.D from Rayalaseema University, Kurnool in 2017 in Digital Image processing. He has total 26 years of teaching experience and his area of interesting is Machine learning, Deep learning and Image processing. Currently

he is working as a professor in Dept of CSE, PBR VITS, KAVALI.

Ch. Lekhya Harshika , B.Tech with Specialization of Computer Science and Engineering-Artificial Intelligence in PBR Visvodaya Institute of Technology and Science, Kavali.

T. Veera Ashrith Mehar , B.Tech with Specialization of Computer Science and Engineering-Artificial Intelligence in PBR Visvodaya Institute of Technology and Science, Kavali.

Ch. Lakshmi , B.Tech with Specialization of Computer Science and Engineering-Artificial Intelligence in PBR Visvodaya Institute of Technology and Science, Kavali.

G. Yaswin Sai Goud , B.Tech with Specialization of Computer Science and Engineering-Artificial Intelligence in PBR Visvodaya Institute of Technology and Science, Kavali.