IJMTST is UGC Approved Journal with Journal Id: 43137

International Journal for Modern Trends in Science and Technology, 8(S08):52-56, 2022

ISSN: 2455-3778 online
DOI: https://doi.org/10.46501/IJMTST08S0810
Available online at: https://www.ijmtst.com/vol8si08.html

Practical Speech Emotion Recognition

Jajjara Bhargav[1] | Tata Venkateswarlu[2] | P Susmitha Vadana[3]

[1]Computer Science and Information Technology, Chalapathi Institute of Engineering and Technology, Guntur, AP, India.
[2]Computer Science and Engineering, Sri Mittapalli College of engineering, Guntur, AP, India.
[3]Computer Science and Engineering, Tirumala Engineering College, Guntur, AP, India.

To Cite this Article
Jajjara Bhargav, Tata Venkateswarlu and P Susmitha Vadana. Practical Speech Emotion Recognition. International Journal for Modern Trends in Science and Technology 2022, 8(S08), pp. 52-56. https://doi.org/10.46501/IJMTST08S0810

Article Info
Received: 26 May 2022; Accepted: 24 June 2022; Published: 28 June 2022.

ABSTRACT

***SPEECH EMOTION RECOGNITION** is where emotions can be recognized from the speech. Speech is the most normal way to express yourself as human beings. Extending this means of communication to computer applications is only inevitable then. It describes speech emotion recognition (SER) systems as a set of methodologies that process and classify voice signals to detect the emotions embedded. It is used an MLP Classifier for this and made use of the sound file library to read the sound file, and the librosa library to extract features from it. Since emotions help us understand each other better, applying this understanding to computers is a natural outcome. Thanks to the smart mobile devices that are able to recognize and respond to voice commands with synthesized speech, speech recognition is already in our everyday lives. Recognition of speech emotions (SER) may also be used to enable them to detect our emotions.*

KEYWORDS: Fear, anger,sadness,joy,Disgust

1. INTRODUCTION

Speech Emotion Recognition is software used to recognize the emotions of humans. Attributes of human voice such as pitch, timbre, loudness and tone make human voice versatile for communication. It can be observed that humans can convey their emotions, even by changing the specified characteristics. This helps the human emotion to be defined by speech analysis. Speech Emotion Recognition recognizes the various emotions like happy, sad, anger, and many more.

➢Fear: emotion comes with an unpleasant situation caused from pain, Anger or feeling afraid.

➢**Anġer:** involves a strong feeling of aggravation, uncomfortable situation stress, displeasure, or hospitality.

➢**Sadness:** A feeling caused with disadvantage or loss due to anything.

➢**Joy:** feeling happy. Other words are happiness, gladness.

➢**Disgust:** A feeling with strong disapproval, nasty, dislike

➢**Surprise:** occurred with an unexpected event or shock. Picture (with "Float over text" unchecked). The tonal quality not only changes with different emotions and moods but the associated patterns of speech also shift. For example, when they are angry, people may tend to speak loudly and use shrill or high pitched voices while they are in an emotional state of fear or panic. Many people are likely to ramble when they are nervous or excited. Sound speech characteristics should be used in cases where face to face communication is not possible or where there is no readily accessible language

constraint and proper model for lexicon based speech analysis.

Following are many conditions where speech characteristics can be used as a means of classifying human emotions:

i. Play music and change ambient room lighting to the sound of the conversation.

ii. Carrying out social science work

iii. Customer service centers may gain insight into customer loyalty by actually.

The main objective of employing speech emotion recognition is to adapt the system response upon detecting frustration or annoyance in the speaker's voice. So, it helps to know the state of speaker's emotion and results in best experience to the users.

2. PROPOSED METHODOLOGY

Speech Emotion Recognition, abbreviated as SER. Speech is a complex signal, It contains the information regarding the message, the speaker, the language and the motions. Emotion makes speech more attractive, more effective and more expressive. So, Speech Emotion Recognition (SER) means to understand emotional state of a human by extracting features from his/her voice.

Speech Emotion Recognition contains five main modules as seen in figure 1 below.

1. Speech input
2. Feature Extraction
3. Feature Selection
4. Classification
5. Recognized Emotion

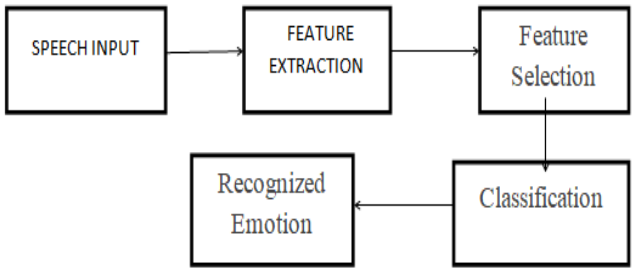


Figure 1: Modules in SER

3. PROPOSED METHODOLOGY

An extraction of these speech features which represents emotions is an important factor in speech emotion recognition system. Feature extraction means extracting the desirable features to extract emotion of human from the speech.

There are mainly two types of features.

Prosodic Features

1. Specral Features
2. Prosodic Features:

Prosodic features are never deals with what we speak, but it deals with how we speak i.e., the loudness, the pitch, the energy, the stress or the rhythm given to speech we dealing.

Spectral Features:

Spectral features are frequency based features. Mel Frequency Cestrum Coefficient (MFCC), Linear Predictive Coding (LPC), Perceptual Acoustic Features (PLP) is the some popular techniques to determine or to extract features from the speech.

4. MODULES USED

DEEP LEARNING MODELS

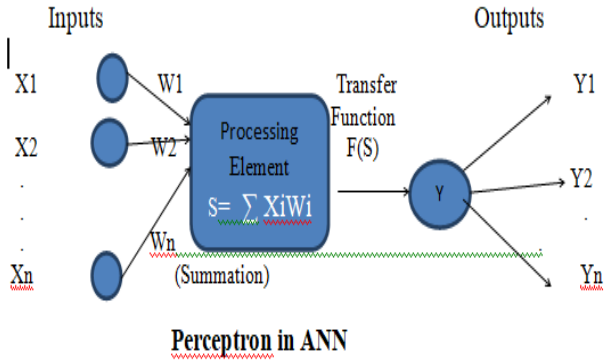
This section is about what is deep learning and Artificial Neural Networks which performs both Feature Extraction and Classification and some popular training models that are used for Feature Extraction.

DEEP LEARNING:

It is an advanced field of Machine Learning that uses the concepts of Neural Networks to solve highly computational use cases that involves the analysis of multidimensional data. The main use of this deep learning is it directly automates the process of feature extraction, classification asking sure that very minimal human intervention is needed.

ARTIFICIAL NEURAL NETWORKS:

Artificial Neural Network is basically a computing system that is designed to simulate the way the human brain analysis and process the information. These are have self learning capabilities, that enable it to produce better results as more data become available. So, if you train the network on more data it will be more accurate. We can configure the neural network for specific applications also i.e., Pattern Recognisation, data Classification etc.



Perceptron in ANN

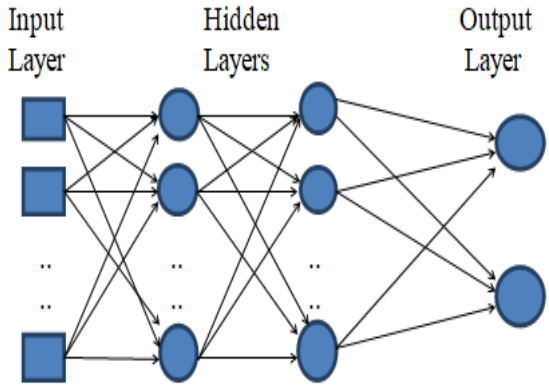


Figure : Multilayer Perception

➢firstly we have multiple inputs X1,X2,.....Xn and we have weightage (X1=W1,X2=W2,....., Xn=Wn).

➢ calculate the weightage sum of these inputs and pass it to Activation Function.

$S = \sum X_i W_i$

➢Activation Function provides Threshold Value.

$S= W^T. X= \sum_{i=1}^{n} W_i X_i$

F(S) = (0 if z<0) or (1 z>=0)

$F(S)= F(W^T.X)$

➢ with that threshold value our output neuron will fire otherwise doesn't fire.

MODES IN PERCEPTRON:

Training Mode:

In the training Mode, the neuron can be trained to fire (or not), for a particular input patterns.

Using Mode:

In the Using mode, when a taught input pattern is detected as the input, its associated output becomes the current output.

Steps to build the MLP classifier:

➢ Initialize the MLP Classifier by defining and initiating the required Parameters.

➢ Data is given to the Neural Network to train it.

➢ the trained network is used to predict the output.

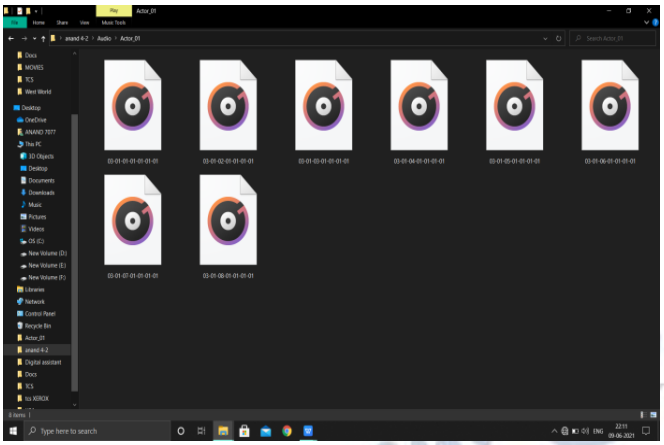
➢ Calculate the accuracy of the predictions

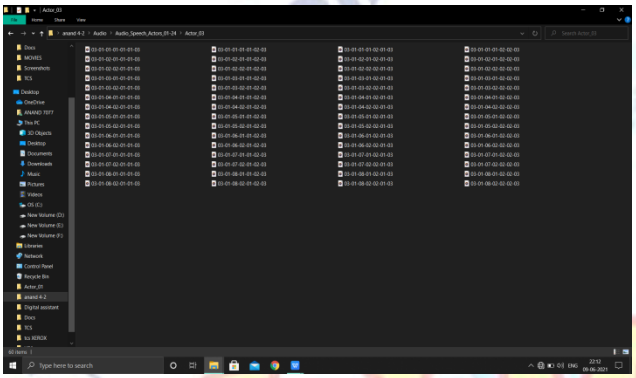
5. DATASETSAND PACKAGES USED

This paper aims to classify different types of emotions such as sad, happy, neutral, angry, disgust, surprised, fearful and calm. In this project, the emotions in speech are predicted using neural networks. This Paper uses Multi Layer Perception (MLP Classifier) for the classification of emotions. (Ryerson Audio-Visual Database of Emotional Speech and Song Dataset) is the dataset used in this project to predict the emotions.

RAVDESS dataset has recordings of 24 actors (i.e., 12 male actors and 12 female actors), the actors are numbered from 01 to 24.The odd numbered are male actors and even numbered are female actors. The emotions contained in the dataset are sad, happy, neutral, angry, disgust, surprised, fearful and calm expressions. Thus, the part of the RAVDESS that contains 60 trials for each of the 24 actors .the we have 1440 files in total. The dataset is labeled in accordance with the decimal encoding and every file has a unique filename. The filename is made up of 7-part numerical identifier. the third numerical part of the filename denotes a label to the corresponding emotion. The emotions are labeled as follows: 01-'neutral', 02-'calm', 03-'happy', 04-'sad', 05-'angry', 06- 'fearful', 07-'disgust', 08 -'surprised'.

Our own Audio Dataset as follows:





PACKAGES USED:

Librosa:

Librosa is a python library for analyzing audio and music. It provides the building blocks necessary to create music information retrieval systems. It has aflutter package layout, standardizes interfaces and names, backwards compatibility, modular functions, and readable code.

Numerical Python (Numpy):

Numpy is a fundamental package use or mathematical and numerical analysis. It is a fast, flexible container for large datasets in python.

Sound file:

Sound file is a python library used to read the sound file.

Glob:

Glob module is used to retrieve the all file paths that match a specific pattern. Sci-kit Learn module is used to build machine learning models. This library contains a lot of efficient tools for machine learning and statistical modeling including classification, regression, and clustering and dimensionality reduction.

Pickle:

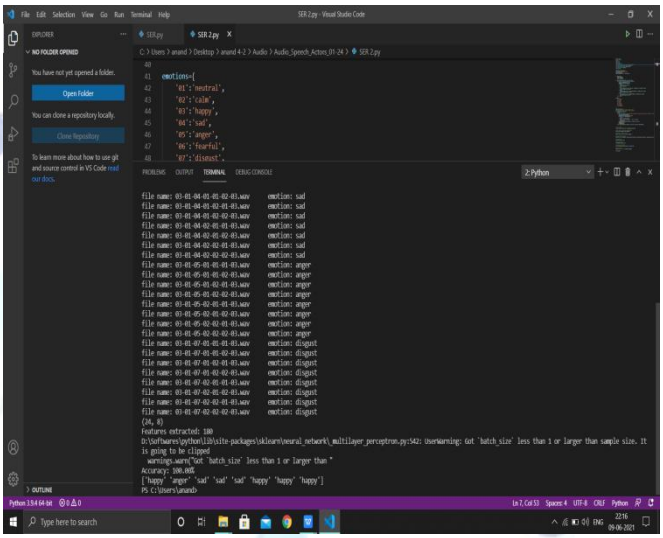
Pickle is a python module used to serialize a python object into a binary Format and desterilize it back to python object.

Pydub:

Pydub package is able to read and save wav file, but we need some type of audio package to actually play sounds.

6. PERFORMANCE ANALYSIS

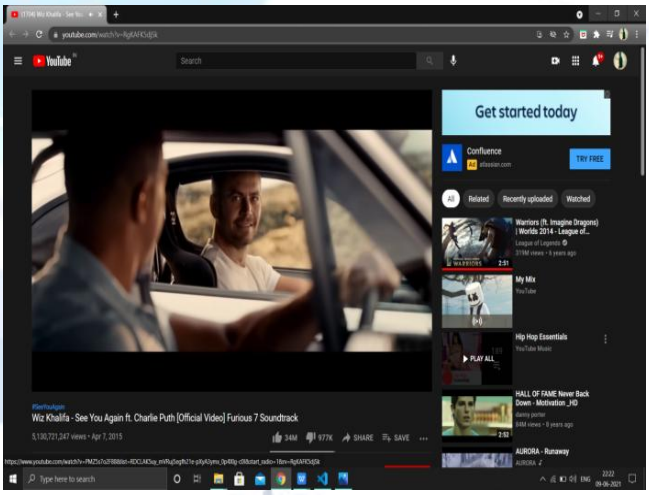
Application 1



Result:

So, finally figure 8.1 shows appropriate emotions of audio files which is present in the corresponding Actor directory with an average accuracy of 85%.

Application 2



Result:

This application finds the emotion of a person and helps to play music according to the person's mood.

7. CONCLUSION

Many databases available for Speech Sentiment Analysis have given rise to emotions. That is, it includes samples of speech formed in a given emotion by the equivalent utterances of a voice. Since these speeches are a

deliberate effort, it may not always be like a more ordinary unprompted voice. The main downside in collecting unprompted speech samples, however, is that more human effort and time will be needed. This will also mean collecting speech samples all the time which may contribute to questions about privacy. Most used methods of feature extraction and MLP classifier performances are reviewed. Success of emotion recognition is dependent on appropriate feature extraction as well as Proper classifier selection from the sample emotional speech. It can be seen that Integration of various features can give the
better recognition rate. Classifier performance is needed to be increased for recognition of speaker independent systems. The application area of emotion recognition from speech is expanding as it opens the new means of communication between human and machine. It is needed to model effective method of speech feature extraction so that it can even provide emotion recognition of real time speech.

Conflict of interest statement

Authors declare that they do not have any conflict of interest.

REFERENCES

[1] S. S. Narayanan, "Toward detecting emotions in spoken dialogs," IEEE Trans.Speech Audio Process., vol. 13,no. 2, pp. 293–303, Mar. 2005.

[2] B. Yang and M. Lugger, "Emotion recognition from speech signals using new harmony features," Signal Processing, vol. 90, no. 5, pp. 1415–1423, May 2010.

[3] H. Cao, R. Verma, and A. Nenkova, "Speaker sensitive emotion recognition via ranking: Studies on acted and spontaneous speech," Comput. Speech Lang., vol.28.1. pp. 186–202, Jan. 2015.

[4] L. Chen, X. Mao, Y. Xue, and L. L. Cheng, "Speech emotion recognition: Features and classification models," Digit. Signal Process. vol. 22, no. 6, pp.1154– 1160, Dec. 2012.

[5] T. L. Nwe, S. W. Foo, and L. C. De Silva, "Speech emotion recognition using hidden Markov models," Speech Common., vol. 41, no. 4, pp. 603–623,Nov.2003.

[6] J. Rong, G. Li, and Y.P. P. Chen, "Acoustic feature selection for automatic emotion recognition from speech," Inf. Process. Manag., vol. 45, no. 3, pp.325-328 May 2009.

[7] E. M. Albornoz, D. H. Milone, and H. L. Rufiner, "Spoken emotion recognition using hierarchical classifiers," Comput. Speech Lang., vol. 25, no. 3, pp.556-570 Jul. 2011.

[8] J.H. Yeh, T.L. Pao, C.Y. Lin, Y.W. Tsai, and Y.T. Chen, "Segment based emotion recognition from continuous Mandarin Chinese speech," Comput. Human Behav., vol. 27, no. 5, pp. 1545–1552, Sep. 2011.

[9] W. Dai, D. Han, Y. Dai, and D. Xu, "Emotion Recognition and AffectiveComputing on Vocal Social Media," Inf. Manag., Feb. 2015.