International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 08 Issue: 07 | July 2021 www.irjet.net p-ISSN: 2395-0072

Speech Emotion Recognition Using Machine Learning

Kumari S[1], Perinban D[2], Balaji M[3], Gopinath D[4], Hariharan S J[5]

[1]*Assistant Professor, Department of Information Technology,Panimalar Engineering College, Anna University, Chennai.*

[2,3,4,5]*Department of Information Technology, Panimalar Engineering College, Anna University, Chennai.*

Abstract - *Speech emotion recognition could even be a challenging task, and extensive reliance has been placed on models that use audio features in building well-performing classifiers. In the lifetime of humans emotions play an important role in communication, the detection and analysis of an equivalent is of important importance in today's digital world of remote communication. Emotion detection may be a challenging task, because emotions are subjective. There is no common consensus on the way to measure or categorize them. We define a speech emotion recognition system as a set of methodologies that process and classify speech signals to detect emotions embedded in them. In this study we decide to detect underlying emotions in recorded speech by analyzing the acoustic features of the audio data of recordings. Emotion is an integral a part of human behavior and inherited property altogether mode of communication. We, human is well trained thought your experience reading recognition of varied emotions which make us more sensible and understandable. But just in case of machine, however, it can easily understand content based information like information in text, audio or video but still far behind to access the depth behind the content. There are three classes of features during a speech namely, the lexical features (the vocabulary used), the visual features (the expressions the speaker makes) and therefore the acoustic features (sound properties like pitch, tone, jitter, etc.).*

Key Words: Emotion recognition, Emotion detection, lexical features, visual features, acoustic features

1. INTRODUCTION

As emotional dialogue consists of sound and spoken content, our model encodes the knowledge from audio and text sequences using dual recurrent neural networks (RNNs) then combines the information from these sources to predict the emotion class. This architecture analyzes speech data from the amplitude to the language level, and it thus utilizes the knowledge within the info more comprehensively than models that specialize in audio features. Extensive experiments are conducted to research the efficacy and properties of the proposed model. Our proposed model can

2.2 Emotion Detection

The speech emotion detection system is implemented as a Machine Learning (ML) model. The steps of implementation are comparable to any other ML project, with additional fine-tuning procedures to make the model outperform previous state-of-the-art methods in assigning data to a minimum of one among 4 emotion categories (i.e., angry, happy, sad and neutral).

1.1 Objective of the Project

Choosing to follow the lexical features would require a transcript of the speech which might further require a further step of text extraction from speech if one wants to predict emotions from real-time audio. Similarly, going forward with analyzing visual features would require the surplus to the video of the conversations which could not be feasible in every case while the analysis on the acoustic features are often wiped out real-time while the conversation is happening as we'd just need the audio data for accomplishing our task. Hence, we elect to analyze the acoustic features during this work. The field of study is termed as Speech Processing and consists of three components: Speaker Identification, Speech Recognition, Speech Emotion Detection.

1.2 Scope of the Project

An emotion one out of a delegated set of emotions is identified with each unit of language (word or phrase or utterance) that was spoken, with the precise start of every such unit determined within the continual acoustic signal. A striking nature unique to humans is that the ability to change conversations supported the spirit of the speaker and also the listener.

2. LITERATURE SURVEY

2.1 Emotion Recognition

With the advancement of deep learning methods, more complex neural based architectures are proposed. These neural network-based models are combined to produce higher-complexity models and these models achieved the best-recorded performance when applied to the IEMOCAP dataset. One researcher utilized the multi object learning approach and used gender and naturalness as auxiliary tasks that the neural network- based model learned more features from different dataset.
function better. The flowchart represents a pictorial overview of the process . The first step is data collection, which is of prime importance. The model being developed will learn from the data provided to it and all the decisions and results that a developed model will produce is guided by the data. The second step, called feature

engineering, is a collection of several machine learning tasks that are executed over the collected data. These procedures address the several data representation and data quality issues.

2.3 Ranking SVM Approach

This proposed is a system that considered that the emotion expressed by humans are mostly a result of mixed feeling. Therefore, they suggested an improvement over the SVM algorithm that would consider mixed signals and choose the most dominant one. For this purpose, a ranking SVM algorithm was chosen. The ranking SVM takes all predictions from individual binary classification SVM classifiers also called as rankers, and applies it to the final multi-class problem. Using the ranking SVM algorithm, an accuracy of 44.40% was achieved in their system.

3. DESIGN OF SYSTEM

3.1 Design of the Proposed System

After clicking on the mic button, it starts to listen. And it will start the pre-processing stage, where it will remove the noises and balance the frequency with the help of pre-emphasis and equalization. After that, the noise removed texts will be compared with the datasets, which is customized by us. And if the text's equivalent found the result will be sent back or else it will display can't predict the emotion. If the word is found, then the equivalent emotion will be displayed as a result in the graphical view manner.

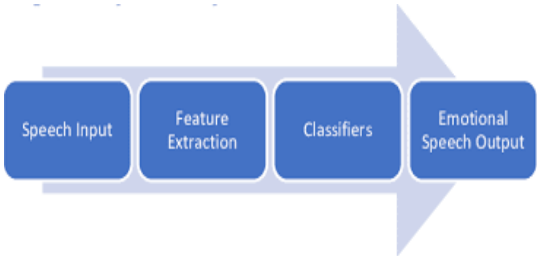

Fig-1: Proposed System Architecture

3.2 List of Modules

1. Voice Input

In this module, the user have to speak up to the mic after pressing the speak button .It will start receiving the user's voice.

2. Voice to Text

In the second module, After receiving the voice, the MFCC, LPCC and PLP Features are performed on the voice to assure the normal hearable frequencies. Then the voice will be converted to text with the help of Google API Speech to Text.

3. Analyzing Texts extracted

In the third module, the results of the previous module Will be i.e. the converted texts are analyzed with the customized datasets.

4. Graphical Result

In the Final module, After comparing the texts with the datasets, a graphical based result will be displayed showing whether the emotion is anger, happy, neutral, etc.

3.3 Speech Processing Module

This is the Module explaining our project includes input speech signal, speech processing, classification and finally detects the Emotion of the input speech.

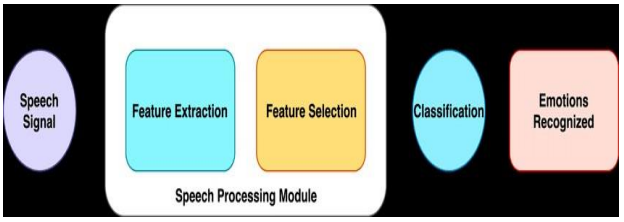

Fig-2: Speech Processing Module

3.4 Pre-Processing Module

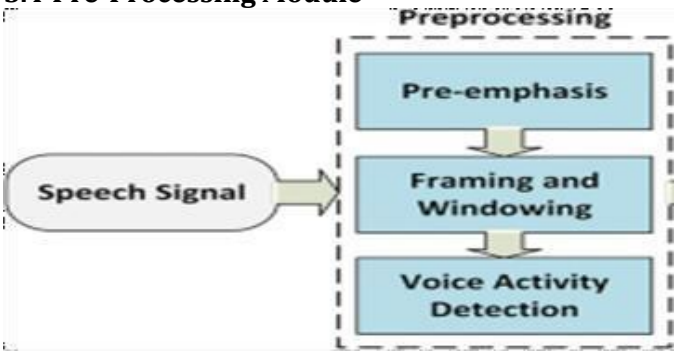

Fig-3: Pre-Processing Module

This is the Pre Processing module, once after getting the input from user, the input speech is preprocessed.

4. CASE STUDY

The case study has like main objective to describe the behavior of this project

4.1 Operation of the System

At the beginning, our system looks like,

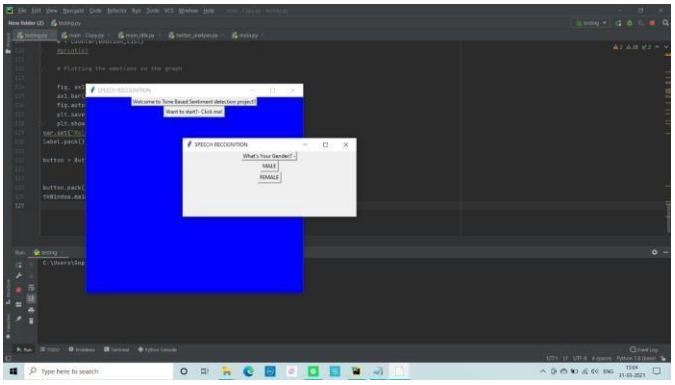

Fig-4: Sample Screen 1

When you select the Gender, it starts listening.

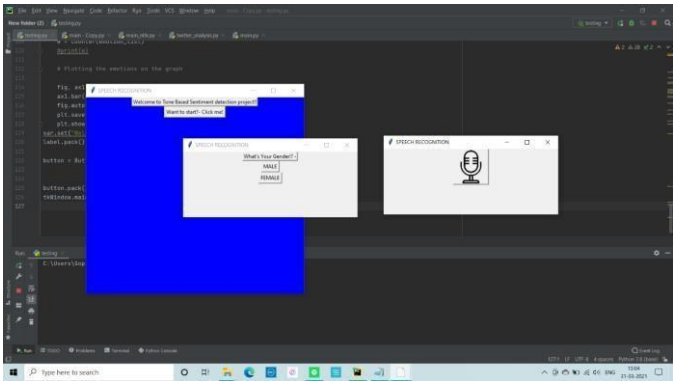

Fig-5 : Sample Screen 2

After clicking on mic, it starts listening.

Fig-6: Listening Screen

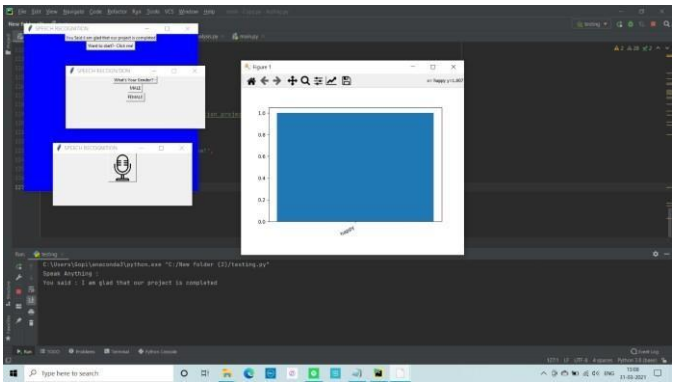

Fig-7: Final Graph

4.2 Result Analysis

The evaluation of the speech emotion recognition 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.

5. CONCLUSIONS

For good emotion recognition system mainly three databases are used. On the basis of ability, they have to recognize a speech recognition system can be separated in different classes are isolated, connected, spontaneous and continuous words. Relevant emotional features extraction from the speech is the second important step in emotions recognition. To classify features there is no unique way but preferably acoustic and linguistic features taxonomy is considered separately.

There are a number of methods for feature extraction like Linear predictive cepstral coefficients (LPCC), Power spectral analysis (FFT), First order derivative (DELTA), Linear predictive analysis (LPC), Mel scale cepstral analysis (MEL), perceptual linear predictive coefficients (PLP) and Relative spectra filtering of log domain coefficients (RASTA) and some of them are briefly covered in this paper. Another important part of speech emotion recognition system is the use of classifier. In the paper, the detailed review on KNN, SVM, CNN, Naive Bayes, and recurrent neural network classifier for speech emotion recognition system.

The last section of the paper covers the review on the use of the deep neural network to make speech emotion recognition system. To further improve the efficiency of system combination of more effective features can be used that enhances the accuracy of speech emotion recognition system. Thus this concludes the SER system.

REFERENCES

[1] M. E. Ayadi, M. S. Kamel, F. Karray, —Survey on Speech Emotion Recognition: Features, Classification Schemes, and Databases‖, Pattern Recognition, vol. 44, pp. 572-587, 2011.

[2] S. K. Bhakre, A. Bang, —Emotion Recognition on The Basis of Audio Signal Using Naive Bayes Classifier‖, 2016 Intl. Conference on Advances in Computing, Communications and Informatics (ICACCI), pp. 2363- 2367, 2016.

[3] I. Chiriacescu, —Automatic Emotion Analysis Based On Speech‖, M.Sc. THESIS Delft University of Technology, 2009.

[4] X. Chen, W. Han, H. Ruan, J. Liu, H. Li, D. Jiang, —Sequence-to-sequence Modeling for Categorical Speech Emotion Recognition Using Recurrent Neural Network‖, 2018 First Asian Conference on Affective Computing and Intelligent Interaction (ACII Asia), pp. 1-6, 2018.

[5] P. Cunningham, J. Loughrey, —Over fitting in WrapperBased Feature Subset Selection: The Harder You Try the Worse it Gets Research and development in intelligent systems‖, XXI, 33-43, 2005.

[6] C. O. Dumitru, I. Gavat, —A Comparative Study of Feature Extraction Methods Applied to Continuous Speech Recognition in Romanian Language‖, International Symposium ELMAR, Zadar, Croatia, 2006.

[7] S. Emerich, E. Lupu, A. Apatean, —Emotions Recognitions by Speech and Facial Expressions Analysis‖, 17th European Signal Processing Conference, 2009.

[8] R. Elbarougy, M. Akagi, —Cross-lingual speech emotion recognition system based on a three-layer model for human perception‖, 2013 AsiaPacific Signal and Information Processing Association Annual Summit and Conference, pp. 1–10, 2013.

[9] D. J. France, R. G. Shiavi, —Acoustical properties of speech as indicators of depression and suicidal risk‖, IEEE Transactions on Biomedical Engineering, pp. 829–837, 2000.

[10] P. Harár, R. Burget, M. K. Dutta, —Speech Emotion Recognition with Deep Learning‖, 2017 4th International Conference on Signal Processing and Integrated Networks (SPIN), pp. 137-140, 2017.