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

EMOTION RECOGNITION FROM AUDIO USING LIBROSA AND MLP CLASSIFIER

Prof. Guruprasad G[1], Mr. Sarthik Poojary[2], Ms. Simran Banu[3], Ms. Azmiya Alam[4], Mr. Harshith K R[5]

[1]Assistant Professor , Dept. of CSE, Yenepoya Institute of Technology, Moodbidri, India-574225
[2,3,4,5]Students, Dept. of CSE, Yenepoya Institute of Technology, Moodbidri, India-574225

Abstract - *Emotion detection has become one of biggest marketting strategies in which mood of consumer plays an important role. So to detect current emotion of person and suggest the appropriate product or help him accordingly, the demand of the product will be increased or the company.The Emotion detection is natural for humans but it is very difficult task for machines. In today's world detecting emotions is one of the most important marketting strategy. For this purpose we decided to do a project in which we could detect a person's emotion just by their voice Using Deep and Convolutional Neural Networks for Accurate Emotion Classification on DEAP Dataset.*

***Key Words*: CNN,Deep learning.**

1.INTRODUCTION

Speech emotion recognition(SER) is mostly beneficial for applications, which need human-computer interaction such as speech synthesis, customer service, education, forensics and medical analysis. Speech being a primary medium to pass information ,we humans can also understand the intensity and mood of the speaker by the speech data generated . Recognizing of emotional conditions in speech signals are so challengeable area for several reasons. First issue of all speech emotion methods is select the best features, which will be powerful enough to distinguish between different emotions. The presence of various language, accent, sentences, speaking style, speakers also add another difficulty because these characteristics directly change most of the extracted features includes pitch, energy, etc. Speech emotion recognition is tough because emotions are subjective and annotating audio is challenging. The idea of creating this project was to build a machine learning model that could detect emotions from speech we have with us all the time. In this we have used librosa and MLP classifier, here librosa is used for analyzing audio and music. It has flatter package layout, standardizes interfaces and names, backwards compatibility, modular functions, and readable code. The MLP-Classifier is used to classify the emotions from the given wave of learning rate to be adaptive. In this study we attempt to dettect underlying emotions in recorded speech by analysing the acoustic features of the audio data of recordings. There are three classes of features in speech namely, lexical , visual and acoustic features. The problem of speech emotion recognition can be solved by analysing one or more of these features.

2. LITERATURE SURVEY

[1] Title: "A comparison of the discrete and dimensional models of the emotion in music"

Author: TuomasEerola and Jonna K. Vuoskoski, Psychology of Music, 1-32, The Author(s) 2010.

The primary aim of the present study is to contribute to the theoretical debate currently occupying music and emotion research by systematically comparing evaluations of perceived emotions using two different theoretical frameworks: the discrete emotion model, and dimensional model of affect. The importance of the comparison lies not only in the prevalence of these models in music and emotion studies, but also in the suggested neurological differences involved in emotion categorization and the evaluation of emotion dimensions, aswell as in the categorically constrained affect space the excerpts have represented to date. Moreover, the various alternative formulations of the dimensional model have not been investigated in music and emotion studies before. A secondary aim is to introduce a new, improved set of stimuli – consisting of unfamiliar, thoroughly tested and validated non-synthetic music excerpts – for the study of musicmediated emotions. Moreover, this set of stimuli should not only include the best examples of target emotions but also moderate examples that permit the study of more subtle variations in emotion.

[2] Title: DEAP: A Database for Emotion Analysis using Physiological Signals

Author: Sander Koelstra, *Student Member, IEEE,* Christian M¨uhl, Mohammad Soleymani, *Student Member, IEEE,*Jong-Seok Lee, *Member, IEEE,* Ashkan Yazdani, Touradj Ebrahimi, *Member, IEEE,*Thierry Pun, *Member, IEEE,* Anton Nijholt, *Member, IEEE,* Ioannis Patras, *Member,IEEE.*

In this work, we have presented a database for the analysis of spontaneous emotions. The database contains physiological signals of 32 participants, where each participant watched and rated their emotional response to

40 music videos along the scales of arousal, valence, and dominance, as well as their liking of and familiarity with the videos. We presented a novel semiautomatic stimuli selection method using affective tags, which was validated by an analysis of the ratings participants gave during the experiment. Significant correlates were found between the participant ratings and EEG frequencies. Single-trial classification was performed for the scales of arousal, valence and liking using features extracted from the EEG, peripheral and MCA modalities. The results were shown to be significantly better than random classification. Finally, decision fusion of these results yielded a modest increase in the performance, indicating at least some complementarity to the modalities.The database is made publicly available and it is our hope that other researchers will try their methods and algorithms on this highly challenging database.

3. METHODOLOGY

In this the emotions in the speech are predicted using neural networks. Multi-Layer Perceptron Classifier (MLP Classifier) and RAVDESS(Rayerson Audio-Visual Database of Emotional Speech and Song dataset) are used.

a)Database Description

RAVDESS dataset has recordings of 24 actors, 12 male actors, and 12 female actors, the actors are numbered from 01 to 24. The male actors are odd in number and female actors are even in number. The emotions contained in the dataset are as sad, happy, neural, angry, disgust, surprised, fearful and calm expressions. The dataset contains all expressions in three formats, those are: Only Audio, Audio-Video and Only Video. Since our focus is on recognise emotions from speech, this model is trained on Audio-only data.

A Multi-Layer perceptron(MLP) is a network made up of perceptron. It has an input layer that receives the input signal, an output layer that makes predictions or decisions for a given input, and the layers present in between the input and output layer is called hidden layer. In the proposed methodology for Speech Emotion Recognition, the MLP network will have one input layer, 300,40,80,40 hidden layers and one output layer. The hidden layers will be large numbers and number of hidden layers can be changed as per requirements.

b)Features

The data was acquired directly from the group of Audio files and they were transformed in 264 vectors of features. A wide range of possibilities exist for parametrically representing a speech signal and its content in a vector, with intention to extract a relevant information from it.

The learning process covers two steps as shown in Figure 1, the first step is a forward processing of input data by the neurons that producesa forecasted output, the second step is the adjustment of weights within neuron layers, in order to minimize the errors of forecasted solution compared with the correct output.

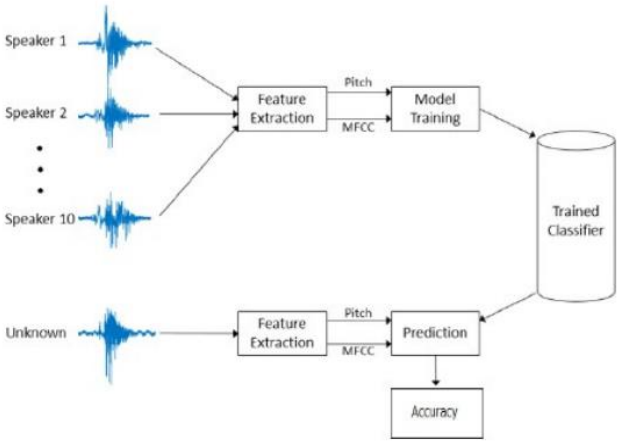


Figure 1. Speech Emotion Recognition System

SYSTEM DESIGN

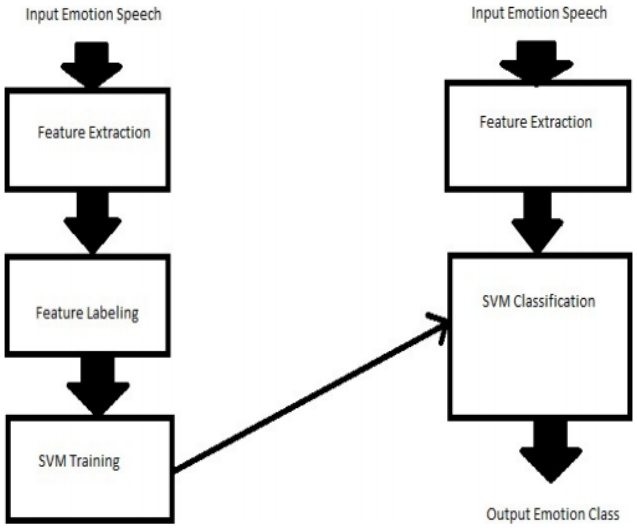
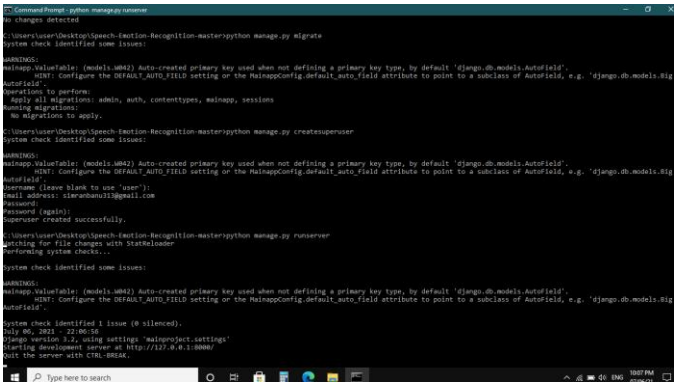


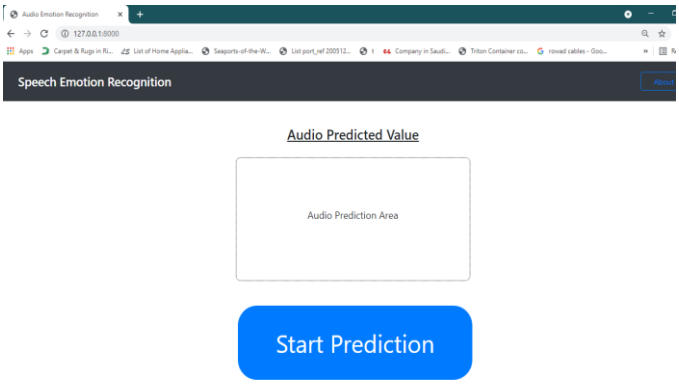
Figure 2. Block Diagram of the System

The model is trained using several user speech input made by different actors. The audio data is analyzed and then its feature is extracted using Librosa library. In the above diagram the flow is shown. After extracting the features its labeled with appropriate label. The trained model is saved using pickle library and then loaded whenever its required. The prediction of emotion is done by taking the audio data from the microphone and preprocessing it i.e., extracting its features and feeding it to the model.The model willpredict the emotion output as it takes the input as shown in Figure 2.

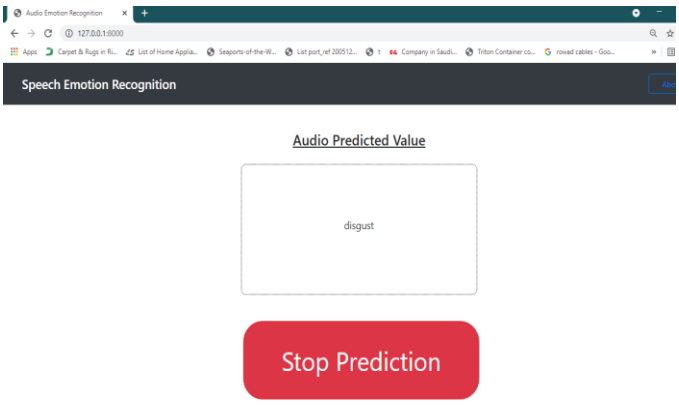
RESULTS



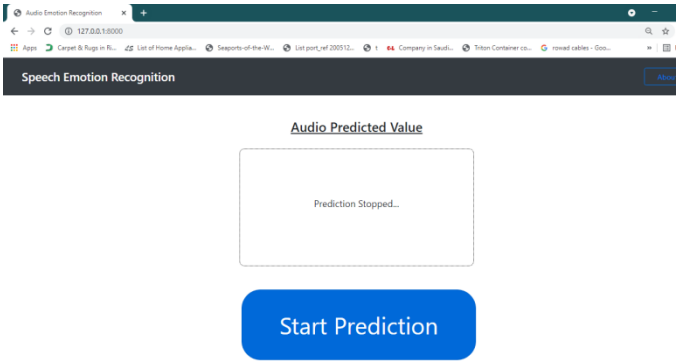
In this page superuser is created so its starts development of server.



By clicking to the start prediction button the voice is detected from the audio to predict the emotions.



Emotion is Rcognized from the audio and detects emotion which will be displayed on the screen by extracting features such as tone, pitch, etc.



Predicts the value and Prediction Stopped.

4. CONCLUSIONS

The overall system results how we can leverage Machine learning to obtain the ultimate emotions from audio data and some insights on the human expression of emotion through voice. This systems can be employed in a variety of setups like Call Centre for complaints or marketting, in voice-based virtual assistants or chatbots, in linguistic research, etc. Some of the possible steps that can be implemented to make the models strongly formed and accurate are the following:

- An accurate implementation of the speed of the speaking can be explored to check if there is any error so it can resolve some of the deficiencies of the model.
- Figuring out a way to clear aimless silence from the audio clip.
- Exploring other acoustic properties of sound data to check their applicability in domain of speech emotion recognition.
- Following lexical features based approach towards SER and using an ensemble of the lexical and acoustic models, will improve accuracy of the system .

REFERENCES

- [1] Tuomas Eerola and Jonna K. Vuoskoski, “A comparison of the discrete and dimensional models of emotion in music”, Psychology of Music, 1–32, The Author(s) 2010.
- [2] Eddie Harmon-Jones, Cindy Harmon-Jones, and Elizabeth Summerell, “On the Importance of Both Dimensional and Discrete Models of Emotion” School of Psychology, The University of New South Wales, Sydney 2052, Australia.
- [3] Roddy Cowie, Ellen Douglas-Cowie, Susie Savvidou, Edelle McMahon, Martin Sawey & Marc Schröder, ‘FEELTRACE’ Schools of Psychology and English, Queen's University Belfast.

- [4] S. Koelstra, C. Muhl, M. Soleymani, J.-S. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, and I. Patras, "Deap: A database for emotion analysis using physiological signals," IEEE Transactions on Affective Computing, vol. 3, no. 1, pp. 18–30, 2012.
- [5] "Using Deep and Convolutional Neural Networks for Accurate Emotion Classification on DEAP Dataset," Proceedings of the Twenty-Ninth AAAI Conference on Innovative Applications(IAAI-17)).
- [6] Chung, Seong Youb, and Hyun Joong Yoon. "Affective classification using Bayesian classifier and supervised learning."Control,Automation and Systems (ICCAS), 2012 12th International Conference on. IEEE, (2012).
- [7] Jaebok Kim, Khiet P. Truong, Gwenn Englebienne, and Vanessa Evers, "Learning spectro-temporal features with 3D CNNs for speech emotion recognition",Human Media Interaction, University of Twente, Enschede, The Netherlands, 2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII).
- [8] Babak Basharirad, and Mohammadreza Moradhaseli,"Speech emotion recognition methods: Literature review",A IP Conference Proceedings 1891, 020105 (2017).
- [9] Kim, Jaebok and Englebienne, Gwen and Truong, Khiet P and Evers, Vanessa, "Towards Speech Emotion Recognition ``in the wild'' using Aggregated Corpora and Deep Multi-Task Learning", 'Proceedings of the INTERSPEECH', year-2017
- [10] B. Yang, M. Lugger, "Psychological motivated multi-stage emotion classification exploiting voice quality feature." F. Mihelic, J. Zibert, Speech Recognition, In-Tech, 2008, chapter 22.
- [11] C.-H. Wu and W.-B. Liang, "Emotion Recognition of Affective Speech Based on Multiple Classifiers Using Acoustic-Prosodic Information and Semantic Labels," IEEE Trans. Affect. Comput., vol. 2, no. 1, pp. 5–20, Jan. 2011.
- [12] Steven R. Livingstone, Frank A. Russo,"The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English",journal.pone.0196391,May 16, 2018
- [13] Monorama Swain, Aurobinda Routray, Prithviraj Kabisatpathy,"Databases, features and classifiers for speech emotion recognition: a review", I. J. Speech Technology 2018.
- [14] S. G. Koolagudi and K. S. Rao, "Emotion recognition from speech: a review," International Journal of Speech Technology, vol. 15, no. 2, pp. 84–115, 2012.