As per UGC guidelines an electronic bar code is provided to seure your paper

International Journal for Modern Trends in Science and Technology, 9(06): 156-160, 2023
Copyright © 2023International Journal for Modern Trends in Science and Technology
ISSN: 2455-3778 online
DOI: https://doi.org/10.46501/IJMTST0906023
Available online at: http://www.ijmtst.com/vol9issue06.html

Detection Of Emotion Based On Voice Using Machine Learning

Balasubbareddy Mallala[1] | D.Rammohanreddy[2]

[1] PG Scholar, Department of CSE, Newton's Institute of Engineering College, Macherla, AP, India
[2]CSE, Newton's Institute of Engineering College, Macherla, AP, India

To Cite this Article
Balasubbareddy Mallala and D.Rammohanreddy. Detection Of Emotion Based On Voice Using Machine Learning. International Journal for Modern Trends in Science and Technology 2023, 9(06), pp. 156-160. https://doi.org/10.46501/IJMTST0906023

Article Info
Received: 26 May 2023; Accepted: 17 June 2023; Published: 18 June 2023.

ABSTRACT

Emotion Detection plays a major role in many domains specifically to satisfy or to improve the customer Experience. It might be possible to include music in call centers whenever one is angry during a call. In addition, a smart car could slow down when one is angry or afraid. Thus, this type of application has much potential in the world to benefit companies and provide safety to users. The present call centers work on 2-way human interaction, based on the situation the service provider must respond. This is not the case every time as the situation may vary. Smart car is not equipped with this model as of now. So here come the emotion detection systems as a solution for many such problems.

We propose a deep learning strategy for emotion recognition based on convolutional neural networks (2D-CNN).In the project the following 5 different emotion classifications are extracted namely calm, happy, sad, angry, fearful.The features are extracted from the audio signal those are MFCC features(Mel-Frequency cepstral coefficients) these describe the nature of the signal. In this model 40 features are extracted from the signal. In code there is the option to enter no.of features. It is optimal if 40 features are given, as studies about the features say that, generally for a audio file we can extract 80 features but in that 40 are spectral features and remaining 40 are temporal features.Temporal features are the features which denote the physical interpretation of the signal like energy of signal.Time-based signals are converted into frequency domain to obtain spectral features. Those are pitches, rhythm.

KEYWORDS: emotion, machine learning, voice, CNN

1. INTRODUCTION

An interesting aspect of human-computer interaction (HCI) is that it incorporates knowledge and architecture from a variety of disciplines.

Emotion recognition is process of identifying human emotions through their facial expressions. Although humans do this automatically, computational methodologies should also be developed so that a system can also recognize emotions in a similar way.

For example, there is a situation when a psychology teacher wants to note the reactions of the student. While doing this it is difficult for him to remember all the expressions and how the student is feeling in every situation. So the best way for him would be to capture all his expressions with a camera and then go through it to analyze his feelings.

Human emotion communication is vital in the relationships between people, according to a 1997 study called Affective Computing. As psychologist Mehrabian puts it in his famous 7%-38%-55% rule, 7% of an impression is created by language, 38% by tone and gestures, and 55% by non-verbal factors such as facial expressions and body movements.

Even this would be a difficult task. So a better way would be to develop a software which recognizes all the emotions. This is not only restricted to psychologists, but also can be used in different fields. To develop a software of this type decades of scientific research has been conducted for developing methods for automated emotion recognition. We have gone through different methods and areas like Signal Processing, Machine Learning, Computer Vision and etc. Many ways were used like primarily they used spectrograms to distinguish the emotions. There are many different algorithms used to improve accuracy. Some people even developed their own CNN networks for this problem and tried to solve it. But even they were able to solve some part of the problem, it's not fully solved. Current methods of Speech emotion Recognition can be divided into dynamic and static sequence-based methods.

2. Problem Definition

A. Idea

Using Machine Learning concepts to detect Emotions through voice can change the user experience. Running this module in the background of any gadget changes human life in many ways. We will run the dataset using three classifiers: convolutional neural networks (CNN), Support Vector Machine (SVM) classifier, and MLP Classifier. Based on the accuracy, an efficient classifier will be picked. Further, Emotion can be detected using that classifier.

B. Methodologies

We are going to use the librosa library in Python language which converts theaudio signal to vector form. We will use convolutional neural networks (CNN), support vector machines (SVM), and MLP classifiers to decide on the algorithm.

In our survey we have come to this conclusion: based on the dataset, the accuracy of the algorithm is varying.

C. Proposed System

We propose a deep learning strategy for emotion recognition based on convolutional neural networks, support vector machines (SVM), and MLP classifiers. A key idea is to train the model using only the MFCC, commonly known as the spectrum of a spectrum. Mel-frequency cepstrum (MFC) has been interpreted differently in MFCC, and it has been proven to be state-of-the-art for formalizing sounds for automatic speech recognition tasks. As a consequence of their ability to represent the amplitude spectrum of a sound wave in a compact vector form, MFC coefficients have been widely used.

Usually, a fixed window size is used to divide the audio file into frames to create statistically stationary waves. Normalizing the amplitude spectrum involves reducing the "Mel frequency scale ". The purpose of this operation is to empathize the frequency more meaningfully for a significant reconstruction of the wave as the human auditory system can perceive.

In [1] proposed speech-based emotion reorganization using a Machine learning algorithm. In [2] described Speech-basedhuman emotion recognition using MFCC. In [3] proposed support vector machine algorithm. In [4-6] described different approaches for the detection of speech-based emotions.

A total of 40 features have been extracted from each audio file. Each audio file was converted to a floating-point time series to generate the feature.

3. DESIGN OF THE PROPOSED SYSTEM

An input signal is given from the dataset using and all the features are extracted from the signal and are classified into different emotions (happy, sad, anger, surprised, and neutral) using 2D CNN. This gives the output of the emotion in real time.

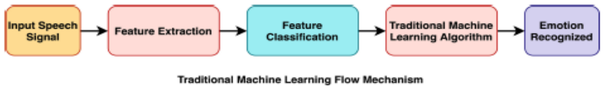
Input Speech Signal
Feature Extraction
Feature Classification
Traditional Machine Learning Algorithm
Emotion Recognized
Traditional Machine Learning Flow Mechanism

Fig.1 Block Diagram

A.Module Description

These are the following modules of our proposed system

i. Input Signal
ii. Feature Extraction
iii. Classification

iv. Applying the Algorithm

v. Output

i. Input Signal

Firstly, we pre-process the given dataset and classify them into different folders accordingto the emotions and assign labels to each emotion. And the input signals are converted into MFCC vector.

The features are extracted from the audio signal those are MFCC features(Mel-Frequency cepstral coefficients) these describe the nature of the signal. In this model, 40 features are extracted from the signal. In the code, there is the option to enter number of features. It is optimal if 40 features are given, as studies about the features say that, generally for a audio file we can extract 80 features but in that 40 are spectral features and remaining 40 are temporal features.

Temporal features are the features which denote the physical interpretation of the signal like energy of signal.

When a time-based signal is converted to a frequency domain signal, spectral features are obtained. Those are pitches, rhythm

ii. Feature Extraction

In this module, the features of the voice files are extracted. We use the concept of Transfer learning where the convolution layer generates features for the voice files. It applies convolution operation on each spectrogram of the voice files and ultimately generates ’ an n-dimensional array which are nothing but learned features of the voice.

At the end of the convolution neural network, we get the bottleneck features of the spectrogram. The bottleneck features are learned from the images and then fed into the MLP, which acts as a top model. In this MLP, the loss function is reduced, and the weights in MLP are updated, as well as the kernels/filters in CNN.

4. IMPLEMENTATION OF THE PROPOSED SYSTEM

An input signal is given from the dataset using and all the features are extracted from the signal and they are classified into different emotions (happy, sad, anger, surprise and neutral) using 2D CNN. This gives the output of the emotion in real time.

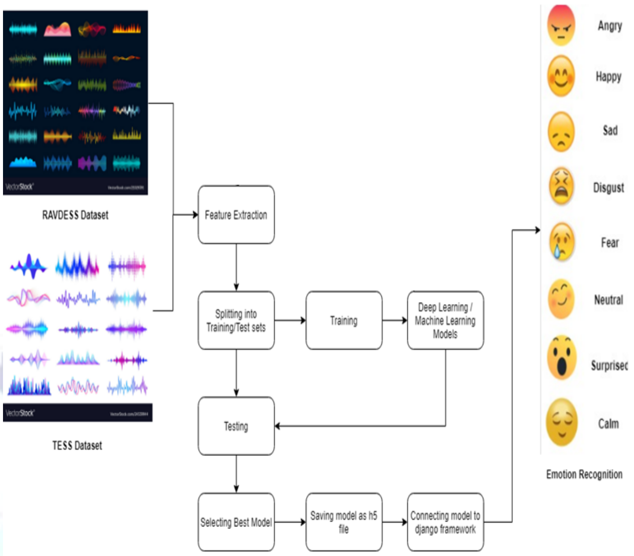

Fig. 2 Flow chart

5. Design and Test steps/criteria

A. Emotion detection

Emotion Detection plays a major role in many domains specifically to satisfy or to improve the customer Experience. It may be possible to include music in call centers when one is angry on the call. When one is angry or afraid, a smart car could slow down. Due to this, this type of application has much potential in the world, allowing companies to benefit as well as providing safety for users.

The present call centers works on 2 way human interaction, based on the situation the service provider must respond. This is not the case every time as the situation may vary. Smart car is not equipped with this model as of now. So here comes the emotion detection systems as a solution for many such problems.

CNN

Using Convolutional Neural Network to recognize emotion from the audio recording. Firstly audio signal is converted into MFCC vector. In MFCC vector there exists various details of the speech like pitch, frequency, amplitude etc.. We use this vector as the input for the CNN model.

Terms involved :

Mel scale — refers to human perception of frequency, a scale of pitches judged equally distant by listeners

Pitch — how high or low a sound is. A higher pitch indicates a higher frequency, whereas a lower pitch indicates a lower frequency

Frequency — measures the frequency of sound vibrations, measured in cycles per second.

Chroma — Representation for audio where spectrum is projected onto 12 bins representing the 12 distinct semitones (or chroma). An octave-by-octave sum of log frequency magnitudes

B. Algorithm

Firstly, the audio files are arranged in the dataset and then it is divided into train set and test set.

Steps involved

Plotting the audio file's waveform and its spectrogram: Audio files are selected from the database and the spectrogram are plotted. Variation in the frequencies can be seen here.

Setting the labels: Labels are set according to the MFCC features which are dominating.

Getting the features of audio files using librosa: In python we have librosa library which converts audio file into the vector form. The vector consists of various fields which judge the emotion of the user.

Changing dimension for CNN model: The dimension of the CNN model is changed according to the input vector and necessary changes are done in the CNN layers.

Removing the whole training part for avoiding unnecessary long epochs list: Audio files consist of many features so for training only few features are used.

Saving the model: Model.h5 file is stored which has the training data and that helps for faster prediction of the emotion.

C. Dataset Description

2452 audio files with 12 male and 12 female speakers, each speaking 2 sentences of equal length in 8 different emotional states, keeping the lexical features (vocabulary) consistent across all speakers.

247 untrained Americans classified speech and song files by eight different emotions at two intensity levels: Calm, Happy, Sad, Angry, Fearful, Disgust, and Surprise, along with a baseline of Neutral for each actor.

Tools:

LibROSA, a Python library, was used to process and extract features from the audio files in this project.

Feature selection is the foundation of modeling. After extracting MFCCs, Chroma, and Mel spectrograms from the audio files we began training machine learning algorithms.

6. RESULTS AND ANALYSIS

After checking with many input files we have observed that based on the number of epochs the accuracy of the model is varying.

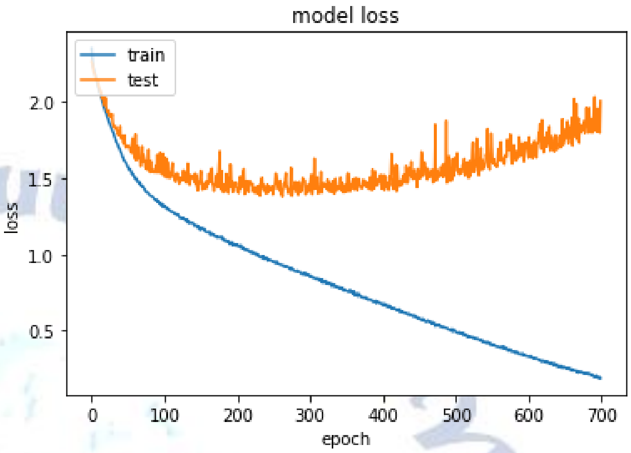

Fig.3 Model Accuracy and Loss Graphs

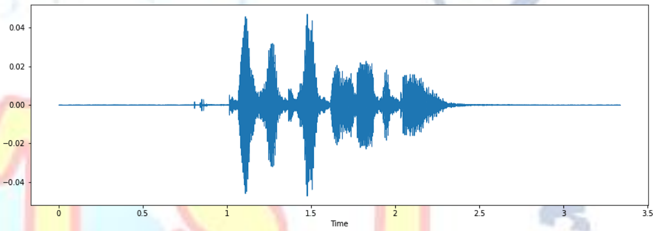

Fig. 4 Wave show function in librosa

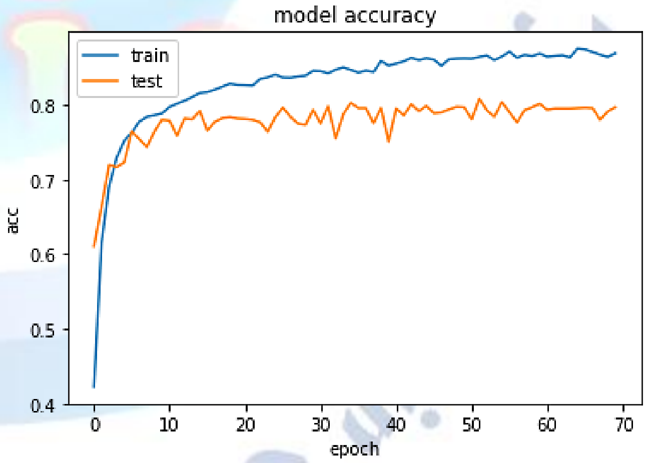

Fig. 5 Model accuracy

7. CONCLUSION

The main objective of "Emotion detection based on voice using machine learning" is to detect the emotion of a person so that necessary changes in the task can be taken. The accuracy of the model which was developed is 80%. This model works very accurately for smaller audio files. So, for larger audio files, it should be converted to smaller ones.

Future Scope

This development of emotion detection through speech will help to improve the customer experience and can provide safety.

- In the future we want to make an app and we want to run this app in the background on mobiles so that each moment emotion can be detected and necessary action can be taken.
- This application can be embedded into any gadget like smartwatches and smart bands. So that at any moment, emotion can be detected and necessary action can be taken.

Conflict of interest statement

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

REFERENCES

[1] Girija Deshmukh, Apurva Gaonkar, Gauri Golwalkar, Sukanya Kulkarn, “Speech based Emotion Recognition using Machine Learning.” (2019), 2019 3rd IEEE International Conference on Computing Methodologies and Communication (ICCMC).

[2] M. S. Likitha, S. R. R. Gupta, K. Hasitha and A. U. Raju, "Speech based humanemotion recognition using MFCC." (2019), EISSN 2227-9709, Published by MDPI.

[3] Sunil Ray, Analytics Vidhya, “Understanding Support Vector Machine algorithms from examples.” (2019).

[4] Naotoshi Seo, "Pitch Detection." (2019). IEEE International Conference on Computing Methodologies 2016.

[5] T. Pao, C. Wang and Y. Li, "A Study on the Search of the Most Discriminative Speech Features in the Speaker Dependent Speech Emotion Recognition."(2019).

[6] SadilChamishka, Ishara Madhavi, RashmikaNawaratne, DammindaAlahakoon, Daswin De Silva,NaveenChilamkurti, Vishaka Nanayakkara, “ A voice-based real-time emotion detection technique using recurrent neural network empowered feature modelling”, Multimedia Tools and Applications (2022) 81:35173–35194 https://doi.org/10.1007/s11042-022-13363-4C. J. Kaufman, Rocky Mountain Research Lab., Boulder, CO, private communication, May 1995.