International Journal of Scientific Research in Computer Science, Engineering and Information Technology
ISSN : 2456-3307 (www.ijsrcseit.com)
doi : https://doi.org/10.32628/CSEIT217446

Speech Emotion Recognition Using MLP Classifier

Nagaraja N Poojary, Dr. Shivakumar G S, Akshath Kumar B.H
Department of Computer Science, Srinivas Institute of Technology Valachil, Mangaluru, Karnataka, India

Article Info
Volume 7, Issue 4
Page Number: 218-222

Publication Issue :
July-August-2021

Article History
Accepted : 10 July 2021
Published : 15 July 2021

ABSTRACT

Language is human's most important communication and speech is basic medium of communication. Emotion plays a crucial role in social interaction. Recognizing the emotion in a speech is important as well as challenging because here we are dealing with human machine interaction. Emotion varies from person to person were same person have different emotions all together has different way express it. When a person express his emotion each will be having different energy, pitch and tone variation are grouped together considering upon different subject. Therefore the speech emotion recognition is a future goal of computer vision. The aim of our project is to develop the smart emotion recognition speech based on the convolutional neural network. Which uses different modules for emotion recognition and the classifier are used to differentiate emotion such as happy sad angry surprise. The machine will convert the human speech signals into waveform and process its routine at last it will display the emotion. The data is speech sample and the characteristics are extracted from the speech sample using librosa package. We are using RAVDESS dataset which are used as an experimental dataset. This study shows that for our dataset all classifiers achieve an accuracy of 68%.

Keywords : Emotion, RAVDESS Dataset, Speech Emotion Recognition, Convolutional neural network.

I. INTRODUCTION

We being a normal people we can easily understand human emotions but it is difficult for a machine to understand. For a machine to understand emotions we use machine learning to teach how it can recognize emotions. Algorithm which build a model by training data in order to make decision or predication without being programmed to do so. Machines are used in multiple area like security, face recognition, email filtering, fruit detection, crop detection and etc. In this project we used convolutional neural network algorithm. CNN takes multiple inputs and find the best output by the required given output. Here we are using a RAVDESS dataset which consist of 1400 audios. Each audio are different from each other and having a different type of emotion in it. We divide the audio by removing the noisy part from the audio and the clean voice is send to the clean folder entire process will be mentioned in further of this document. The audio will go under test and train then at the end display

the emotion. Also we done a live demo section where we give a live audio or a human voice and the machine will predict the correct emotion because it is trained to identify the exact emotion that a voice is having. In future this technology can be used in student teacher emotion interaction so that they can find emotions easily by the use of machine.

II. LITERATURE REVIEW

We design a web end-to-end speech recognition system supported Time-Depth Separable convolutions and Connectionist Temporal Classification. The system has almost 3 times the throughput of a well- tuned hybrid ASR baseline while also having lower latency and a far better word error rate. The process of transcribing speech in real-time from an input audio stream is understood as online speech recognition. The system measure the three a fore mentioned constraints using the following metrics: throughput, Real-Time Factor and latency, and Word Error Rate. This architecture uses Time-Depth Separable convolution as the core building block. Bi-directional RNNs and Transformers either require considerable changes for low-latency deployment or degrade rapidly when limiting the amount of future context [1].

Associative advances in incorporated circuit innovation and PC design have adjusted to make a mechanical domain with for all intents and purposes boundless open doors for development in speech communication applications. In this content, The initial phase in many uses of digital speech processing is to change over the acoustic waveform to a grouping of numbers. Most present day A-to-D converters work by inspecting at an exceptionally high rate, applying digital low pass filter with cutoff set to protect an endorsed data transmission, and after that lessening the sampling rate to the ideal testing rate, which can be as low as double the cutoff frequency of the sharp-cutoff digital filter. This discrete- time representation is the beginning stage for generally applications. Starting here, different representations are gotten by digital processing [2].

The rapid development of technology brings the chance to the reform of English teaching. It is a computer that uses speech processing technology to properly evaluate learners 'pronunciation exercises rather than evaluate by experts. At present, most of the researches on scoring mechanism of CALL system are all about extracting the acoustic characteristics of speech signals, namely, the evaluation of speech segment features, which is obviously not comprehensive enough, ignoring the information hidden in other aspects of speech signals. Therefore, a better scoring method and scoring mechanism should be explored to achieve. In this paper, two speech databases are needed, one is the standard training speech database, which is the standard acoustic model used for training, the other is the speech library to be tested, and it is used for testing and grading. This template library needs to be trained by a large number of samples. In the experiments the HMM model supported template training is employed , because the implicit Maerkekefu model may be a statistical model supported the technology which may well describe the variability and stability of speech signals [3]. The performance degradation of speech applications such as voice search or conversational-bots in noisy and reverberant environment demands the need for improved robustness in automatic speech recognition systems. While several advancements have been made in the acoustic modeling for ASR, the presence of extrinsic noise sources and reverberations continue to pose challenge to the ASR system deployment. The noise robustness are often partly addressed by multi-condition training. In spite of this training, the performance difference between multi-condition train- test and therefore the clean train-test of ASR is pronounced, which warrants the necessity for attaining noise robustness either at speech

representation stage or the training stage. This work focuses on the robust representation learning using unsupervised generative modeling method. The VAE differs from a typical AE where the VAE model assumes that the samples of latent representation are often drawn from a typical Gaussian distribution. We train the CVAE in multi-condition fashion with a small number of filters this way, the model is constrained to primarily learn the speech distribution while ignoring the noise distribution [4].

The purpose of this paper is to style an efficient recurrent neural network based speech recognition system using software with long STM. The design process involves speech acquisition, pre-processing, feature extraction, training and pattern recognition tasks for a spoken sentence recognition system using LSTMRNN. There are five layers namely, an input layer, a totally connected layer, a hidden LSTM layer, SoftMax layer and a sequential output layer. A vocabulary of 80 words which constitute 20 sentences is employed. The depth of the layer is chosen as 20, 42 and 60 and therefore the accuracy of every system is decided. The results reveal that the utmost accuracy of 89% is achieved when the depth of the hidden layer is 42. Since the depth of the hidden layer is fixed for a task, increased performance can be achieved by increasing the number of hidden layers. Speech processing has continuously evolved over the years. State of the art systems is continuously replaced over time. In general, speech processing involves the following: a recognizer or a speech-to-text module that converts speech signals into text, a parser that extracts the semantic context, a dialog manager that determines system response in machine language, an answer generator that provides the system response in text and a speech synthesizer that converts text to the speech signal [5].

II. METHODOLOGY

System implementation is stage where models are converted into a working system, entirely new application is built by replacing the old one using all the freshly implemented design. Dataset, feature extraction, testing and training are the important stages of this project.

A. Proposed System

In this project, we are recognizing emotions from a speech. We used an MLP Classifier for this project and used a sound file library to read the sound file, and the librosa library to extract features. So this will lead us to a better accuracy of detecting the emotion of human.

B. Dataset

For this project we are using the RAVDESS dataset, this is the audio-visual database of emotion speech. This data base has 4300 audio files which includes 4 different emotions which is collected from 10 different actors

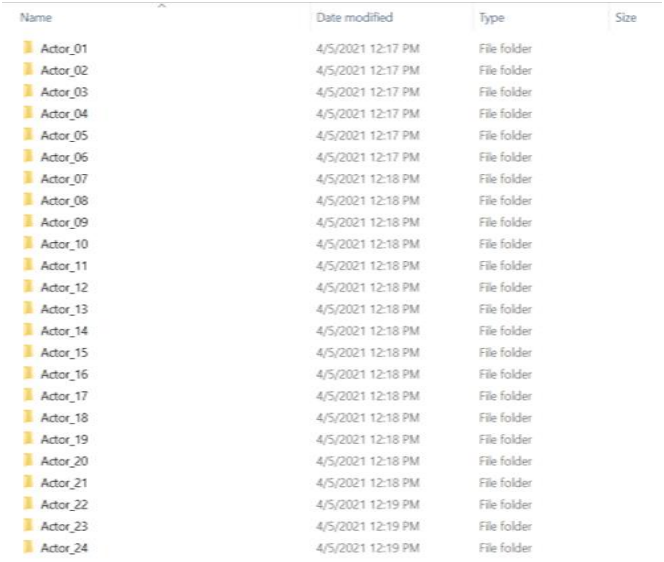

Fig 1. Folder containing audio files

There are 8 emotions basically but here we are dealing with only 4 that are happy, sad, surprised, and angry.

```
emotions={
  '01':'neutral',
  '02':'calm',
  '03':'happy',
  '04':'sad',
  '05':'angry',
  '06':'fearful',
  '07':'disgust',
  '08':'surprised'
}
#These are the emotions User wants to observe more :
observed_emotions=['sad', 'happy', 'surprised','angry']
```

Fig 2. Emotion in dataset

C. Feature extraction

It is the function where we are extract the mfcc, chroma, and mel features from a sound file. This will take 4 parameters, the file name and three Boolean parameters.

The three features are mfcc : mel frequency cepstral coefficient, represent the power spectrum of a sound.

Chroma: pertains to the 12 pitch classes mel: mel spectrum frequency Sound file will be opened and it will be readied and result will be saved to array. For each of three, if it exists then a call will be made to the corresponding function from librosa. Mean value will be noted and result along with feature value and storing it in a file.

D. Training and Testing

We are loading the data where it takes in the relative size of the test set as parameter. X and Y are empty lists, functions will checks whether the emotion are in the list of observed emotions. The feature will be send to X and emotions to Y. Now the testing and training function will be called. 75% of audio will be tested at the same time 25% of audio will trained. For classification we are using MLP Classifier.

```
import numpy as np
x_train,x_test,y_train,y_test=load_data(test_size=0.25)
# print(x_train[1].shape,y_train[1])
```

Fig 3. Traing and testing a speech

Finally the system will be ready and it will trained using fit/train model.

```
#Train the model
model.fit(x_train,y_train)
```

```
MLPClassifier(alpha=0.01, batch_size=256, hidden_layer_sizes=(300,),
              learning_rate='adaptive', max_iter=500)
```

Fig 4. Training the model

III.RESULT

The speech emotion recognition of human voice using machine learning is successfully obtained. By using just small amount of dataset, 69% accuracy has been obtained. If more dataset are used then accuracy will also increase. It is seen that speech emotion recognition using MLP is very efficient and easy to implement.

IV.CONCLUSION

Various experiments are conducted by changing the several parameters like dimension of the model, number of epochs, changing the partition ratio between training and test data set. Different accuracy was found for different experiments. A lighter CNN architecture with 75% of training data and 25% of test data gave good result compared to deeper CNN architecture while classifying among ten classes. The accuracy of this model was found to be 69%. The performance of deeper CNN model was found to be excellent when the classification was among two classes the rationale is there have been a greater number of coaching samples available to classify among two classes. When an equivalent model was wont to classify among ten classes the training dataset was divided into ten labels this led to the less number of coaching samples available for each class. This led us to the poor accuracy of the model. With the greater number of coaching samples available for every class and with the assistance of GPU's to hurry up the training process more accuracy are often achieved in the future enhancements.

V. FUTURE SCOPE

The project can be extended to integrate with the robot to help it to have a better understanding of the mood the corresponding human is in, which will help it to have a better conversation as well as it can be integrated with various music applications to recommend songs to its users according to his/her emotions, it can also be used in various online shopping applications such as Amazon to improve the product recommendation for its users. Moreover, in the upcoming years we can construct a sequence to sequence model to create voice having different emotions. E.g. sad voice, an excited one etc.

VI. REFERENCES

[1]. Awni Hannun, Ann Lee, Qjantong Xu and Ronan Collobert, Sequence to sequence speech recognition with time-depth deperable convolutions, interspeech 2019, Sep 2019.

[2]. Lawrence R Rabiner Ronald W Schafer, “Introduction to Digital Speech Processing", Vol. 1, Nos. 1–2 (2007) 1–194, 2007 L. R. Rabiner and R. W... Schafer.

[3]. Li, J., Deng, L., Gong, Y. (2014). An Overview of Noise-Robust Automatic Speech Recognition, IEEE/ACM Transactions on Audio Speech & Language Processing, Vol.22, No.4, pp.745-777.

[4]. Jinyu Li, Li Deng, Yifan Gong, and Reinhold Hach- Umbach “An overview of noiserobust sutomatic speech recognition” IEEE/ACM Transactions on Audio, Speech and Language Processing, Vol 22, no 4, pp 745-777, 2014.

[5]. Chang, A X., Martini, B and Culurciello E (2015) ‘Recurrent Neural Networks hardware implementation on FPGA’, arXiv preprint arXiv: 1511.05552.

Cite this article as :

Nagaraja N Poojary, Dr. Shivakumar G S, Akshath Kumar B.H, "Speech Emotion Recognition Using MLP Classifier", International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), ISSN : 2456-3307, Volume 7, Issue 4, pp.218-222, July-August-2021. Available at
doi : https://doi.org/10.32628/CSEIT217446
Journal URL : https://ijsrcseit.com/CSEIT217446