International Journal of Research in Engineering, Science and Management
Volume 4, Issue 7, July 2021
https://www.ijresm.com | ISSN (Online): 2581-5792

Speech Emotion Recognition using Machine Learning

Ashwin V. Gatty[1*], G. S. Shivakumar[2], Kiran Shetty[3]
[1,3]*Student, Department of Computer Science and Engineering, Srinivas Institute of Technology, Mangalore, India*
[2]*Professor, Department of Computer Science and Engineering, Srinivas Institute of Technology, Mangalore, India*

***Abstract*: Language is a basic need for the humans to communicate and speech for its primary medium. Spoken interaction in both between human interlocutors and between humans and machines is inescapably embedded within the laws and conditions of Communication, which comprise the encoding and decoding of meaning as well because the mere transmission of messages over an acoustical channel. Here we deal with this interaction between the human and machine through synthesis and recognition applications. Speech recognition, involves capturing and digitizing the sound waves, converting them to basic language units or phonemes, constructing words from phonemes and contextually analyzing the words to make sure correct spelling for words that sound alike. Speech Recognition is the ability of a computer to recognize the caller's answers to move along the flow of the cell. Emphasis is given on the modeling of speech units and grammar on the basic of hidden markov model and neural networks. Speech recognition allows you to provide input to an application with your voice. The applications and limitations on above subject enlighten the impact of speech processing in our modern technical field.**

***Keywords*: Language, Communication, Speech recognition, Interaction.**

1. Introduction

Speech recognition for sometimes referred to as Automatic Speech Recognition which is the process by which a computer (or other type of machines) identifies spoken words. Basically, it means talking to a computer and having it correctly understand what you are saying by understanding the application to react appropriately or to convert the input speech to another medium of conversation which is further perceivable by another application that can process it properly and provide the user the required result. Speech recognition is an alternate to traditional methods of interacting with a computer, like textual input through a keyboard. An effective system can replace or reduce the reliability on, standard keyboard and mouse input. There are many applications of detecting the emotion of the persons like within the interface with robots, audio surveillance, web-based E-Learning, commercial applications, clinical studies, entertainment, banking, call centers, cardboard systems, computer games, etc.

A. Problem Statement

The problem is to specify the popularity process economically, this involves knowing about the speech signal and to acknowledge so as to specify the processing required by the machine.

B. Existing System

The existing system uses robust speech recognition. The matter is that it's having high latency and low accuracy. It's also time consuming. The speech depends upon the surround environment, length of speech used, quite emotions each individual has. The system is split into two stages, the primary stage may be a training stage where the system first trains for various voices or emotions. The second stage is to acknowledge the pattern by using the characteristics derived from the primary task. Considering the method of speech recognition, within the period of user training or recognition albeit whenever describes one word or one sentences as far as possible within the same way.

C. Proposed System

The project is implemented to recognize the speech of the person by recognizing and hearing the voice of an individual. The system can identify the voice of the individual or a person whether that person is sad, happy or angry. The relevant details about who is predicted to possess those needs, and what features the voice interface has got to meet those anticipated needs.

D. Objective

The objective of Speech emotion recognition using Machine Learning is to acknowledge who is talking, speaker check, where the target is to verify a speakers asserted character from identifying of their speech signal, word spotting, which incorporates observing a speech for the event of determined words. Speech emotion recognition is usually beneficial for applications, which require human-computer interaction like speech synthesis, customer service, education, forensics and medical analysis. Recognizing of emotional conditions in speech signals are so much challengeable area for many several reason.

2. Literature Survey

Peng Song [5] offers Transfer Linear Subspace Learning (TLSL) framework for cross corpus recognition of speech.

*Corresponding author: ashwingatty2014@gmail.com

TLSL approaches, TULSL and TSLSL were taken in count. TLSL aims to extract robust characteristics representations over corpora into the trained estimated subspace. TLSL enhances the currently used transfer learning techniques which only focuses on searching the foremost portable components of characteristics TLSL can reach even better results compared to the 6 baseline techniques with stats significance, and TSLSL gives better outcomes compared to TULSL, actually all the transfer learning is more accurate than usual learning techniques. TLSL significantly excels TLDA, TPCA, TNMF and TCA, the superb transfer learning techniques supported characteristics transformation. A big setback that these early transfer learning methods possess was that they concentrate on searching the portable components of characteristics that tend to ignore less informative section. The less informative parts also are significant when it involves transfer learning results experimented that TLSL is implemented for cross-corpus recognition of speech emotion.

With this paper Jun Deng [6] et al focused on unsupervised learning with automatic encoders of speech emotion recognition. Significantly work was on joining generative and discriminative training, by partially supervised learning algorithms designed to settings where non-labeled data was available. the method had been sequentially evaluated with 5 databases in several settings. The proposed technique enhances recognition performance by learning the prior knowledge from non-labeled data in conditions with a smaller number of libeled examples. These techniques can solve the issues in mismatched settings and incorporate the learnings from different domains into the classifiers, eventually leading to outstanding performance. This shows that the model is having the capacity to form good use of the mixture of labeled and non-labeled data for speech emotion recognition. The residual neural network displayed that intense architectures make the classifier beneficial to tug out complicated structure in image processing.

3. Methodology

In this section we have implemented a live demo of speech which are being recognised by the application.

A. Convolutional neural networks (CNNs)

Convolutional neural networks (CNNs) or shift invariant artificial neural networks (SIANNs) are particular sorts of neural networks that, in their hidden layer they need different filters or regions that answer a selected feature of the input. Their design is predicated on the research by Hubel and Wiesel in 1968, which introduces the visual neural cortex as a spatially specialized structure, during which every region responds to a selected characteristic of the input. One positive perspective of CNNs is that the ability to find out features from high-dimensional input data; however, it also learns features from small variations and distortion appearance that results in the massive storage requirement at the time of development. Hence, in CNNs, there usually exists a layer of convolution followed by a down sampling mechanism. The convolution layer has various filter banks, during which their weights are going to be tuned through often backpropagation, we have introduced a convolutional neural network that's capable of detecting angry, happy, sad, and surprised emotions.

B. Librosa

Librosa may be a python package for music and audio analysis. It provides the building blocks necessary to make music information retrieval systems. It's the start line towards working with audio data at scale for a good range of applications like detecting voice from an individual to finding personal characteristics from an audio. Librosa are often defined as a package which is structured as collection of sub modules which further contains other functions.

C. Feature Extraction

It is the function where we are extract the mfcc, chroma, and mel features from a sound file. This may take 4 parameters, the file name and three Boolean parameters are mfcc, mel, and chroma. Sound file are going to be opened and it'll be readied and result are going to be saved to array. For every of three, if it exists then a call is going to be made to the corresponding function from librosa. Means are going to be noted and result alongside feature value and storing it during a file.

4. Result

We experimented an audio file to urge its characteristics by plotting the waveform.

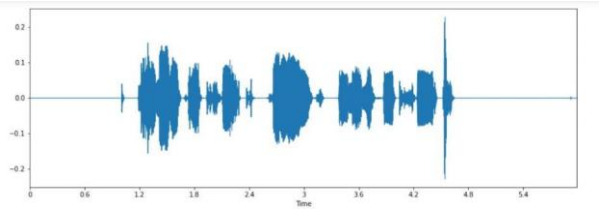

Fig. 1. Plotting a graph for speech signal

We are recording our live audio and saving it in a file.

```
stream = p.open(format=FORMAT,
                channels=CHANNELS,
                rate=RATE,
                input=True,
                frames_per_buffer=CHUNK) #buffer

print("* recording")
frames = []

for i in range(0, int(RATE / CHUNK * RECORD_SECONDS)):
    data = stream.read(CHUNK)
    frames.append(data) # 2 bytes(16 bits) per channel

print("* done recording")
```

Fig. 2. Recording a live audio signal

```
x=[]
ans=[]
file = 'output10.wav'
feature=extract_feature1(file, mfcc=True, chroma=True, mel=True)
x.append(feature)
ans =np.array(x)
# ans=ans.flatten()
# print(len(ans))
# print(ans)
# data.shape
ansr=Emotion_Voice_Detection_Model.predict(ans)
print("This voice is",ansr[0])
```

```
This voice is surprised
```

Fig. 3. Displaying the emotion

The recorded audio is pre-processed and compared with feature extraction and the speech is displayed of the recorded audio.

5. Conclusion

This paper is predicated on speech emotion recognition using convolutional neural network model. During this section, we summarize the datasets, methods, and approaches in SER. Here speech is major source of human understanding so it is often utilized in major events for basic communication. This technology is employed in various field of human machine interaction. Here it is often utilized in online communication were machine understands the human emotion and may ask human at what emotion were human is. We have added tenserflow, mfcc and chroma with the accuracy of quite 68% in future this design is often used. And for more accuracy we've to use more voice data, precision and accent so machine can understand the emotion easily with more accuracy and in future we will develop it for robots which may interact with humans with far more accuracy. In SER system, although there are some progressions in methods and achieved accuracy; however, there exist several restrictions still exist that necessary to be eliminated for a successful recognition system.

6. Future Scope

Our project is often extended to integrate with the robot to assist it to possess a far better understanding of the mood the corresponding human is in, which can help it to possess a far better conversation also because it is often integrated with various music applications to recommend songs to its users consistent with his/her emotions. We even have the choice of mixing a number of the datasets to make a superset. At an equivalent time, this is often possible; there might be problems due to different methods and techniques in creating different databases. As an appropriate solution, we propose exploring the creation of a completely synthetic dataset using generative techniques trained by available datasets. Another challenge which will be addressed in SER is that the difference in emotion expressions in several languages. We believe using transformers, we will build a language-aware model that adapts to the language to classify emotions, and therefore the same concept is often used for different accents during a language.

References

[1] Y. Chen, Z. Lin, X. Zhao, G. Wang and Y. Gu, "Deep Learning-Based Classification of Hyperspectral Data," in IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, vol. 7, no. 6, pp. 2094-2107, June 2014.

[2] L. O. Chua and T. Roska, "The CNN paradigm," in IEEE Transactions on Circuits and Systems I: Fundamental Theory and Applications, vol. 40, no. 3, pp. 147-156, March 1993.

[3] X. Xu, J. Deng, E. Coutinho, C. Wu, L. Zhao and B. W. Schuller, "Connecting Subspace Learning and Extreme Learning Machine in Speech Emotion Recognition," in IEEE Transactions on Multimedia, vol. 21, no. 3, pp. 795-808, March 2019.

[4] Z. Huang, J. Epps, D. Joachim and V. Sethu, "Natural Language Processing Methods for Acoustic and Landmark Event-Based Features in Speech-Based Depression Detection," in IEEE Journal of Selected Topics in Signal Processing, vol. 14, no. 2, pp. 435-448, Feb. 2020.

[5] P. Song, "Transfer Linear Subspace Learning for Cross-Corpus Speech Emotion Recognition," in IEEE Transactions on Affective Computing, vol. 10, no. 2, pp. 265-275, 1 April-June 2019.

[6] J. Deng, X. Xu, Z. Zhang, S. Frühholz and B. Schuller, "Semisupervised Autoencoders for Speech Emotion Recognition," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 26, no. 1, pp. 31-43, Jan. 2018.

[7] M. D. Zeiler et al., "On rectified linear units for speech processing," 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, 2013, pp. 3517-3521.