Graduate Research in Engineering and Technology (GRET)

Volume 1
Issue 6 *Application of Intelligent Computing and Big data Analytics in Healthcare*

Article 7

May 2022

Speech Emotion Recognition System using Librosa for Better Customer Experience

Subhadarshini Mohanty
Odisha University of Technology and Research, Bhubaneswar, Odisha, India, sdmohantycse@cet.edu.in

Subasish Mohapatra
Odisha University of Technology and Research, Bhubaneswar, Odisha, India, smohapatra@cet.edu.in

Amlan Sahoo
Odisha University of Technology and Research, Bhubaneswar, Odisha, India, amlan@cet.edu.in

Recommended Citation

Mohanty, Subhadarshini; Mohapatra, Subasish; and Sahoo, Amlan (2022) "Speech Emotion Recognition System using Librosa for Better Customer Experience," *Graduate Research in Engineering and Technology (GRET)*: Vol. 1: Iss. 6, Article 7.
DOI: 10.47893/GRET.2022.1114
Available at: https://www.interscience.in/gret/vol1/iss6/7

Speech Emotion Recognition System using Librosa for Better Customer Experience

Subhadarshini Mohanty, Subasish Mohapatra, and Amlan Sahoo

Odisha University of Technology and Research, Bhubaneswar, Odisha, India
Email: sdmohantycse@cet.edu.in, smohapatra@cet.edu.in, and amlan@cet.edu.in

Abstract

Call center employees usually depend on instinct to judge a potential customer and how to pitch to them. In this paper, we pitch a more effective way for call center employees to generate more leads and engagement to generate higher revenue by analyzing the speech of the target customer by using machine learning practices and depending on data to make data-driven decisions rather than intuition. Speech Emotion Recognition otherwise known as SER is the demonstration of aspiring to perceive human inclination along with the behavior. Normally voice reflects basic feeling through tone and pitch. According to human behavior, many creatures other than human beings are also synced themselves. In this paper, we have used a python-based library named Librosa for examining music tones and sounds or speeches. In this regard, various libraries are being assembled to build a detection model utilizing an MLP (Multilayer Perceptron) classifier. The classifier will train to perceive feeling from multiple sound records. The whole implementation will be based on an existing Kaggle dataset for speech recognition. The training set will be treated to train the perceptron whereas the test set will showcase the accuracy of the model.

Keywords: Speech Emotion Recognition (SER), Machine Learning, Multi-layer Perceptron (MLP) classifier, Feature extraction, Modular functions, Emotional validity, Mel-frequency cepstral coefficients (MFCCs).

I. INTRODUCTION

The strategy or approach of extricating human feelings from a sent voice signal is defined as Speech Emotion Recognition (SER). This exploits the verifiable truth that an individual's tone and pitch of voice now and again mirror the feeling that the person is encountering. Canines and ponies, for instance, utilize this peculiarity to decipher human inclination. SER is troublesome because speeches are abstract and explaining speech tone is troublesome [1]. Speech examination, then again, offers a significant benefit. They can be used to make savvy frameworks that mix flawlessly into our regular routines as a component of shrewd living. It fills in as the establishment for advancements, for example, voice acknowledgment, voice control, order capacities, and a lot more presented by tech goliaths like Google and Microsoft as Alexa, Cortana, and Samsung's Bixby, as well as other man-made brainpower (AI) applications. Different methods or strategies, like looks, calligraphy investigation, and mental assessments performed regarding the matter, can be utilized to inspire feelings. Notwithstanding this, speech is the most key method for correspondence between any two individuals at some random time [2]. This has incited numerous specialists in an assortment of fields to look at, test, and arrive at positive resolutions in the field of speech examination, bringing about plenty of models and thoughts over the long run. This was well known during the 1950s. With the help of past individuals' information and distributions, the investigators had the option to make a hypothetical method or recipe to interpret speech into a bunch of words for evidently design expectation in an assortment of uses.

This system neglected to create adequate discoveries, and it didn't get the fundamental financing to proceed with the review. Thus, we've chosen to zero in our endeavors regarding this matter. There may be a method to sort out which forecast model to utilize for speech feeling identification arrangement and examination. We utilized librosa and MLP classifier in this, with librosa being utilized for sound and music examination.

II. LITERATURE SURVEY

There are significant works in the literature regarding recognition in speech systems. Learning speech, according to Vincius Maran et al., is a tedious system in which the infant's processing of criteria is highlighted by their randomness on the way to the modern creation of ambient language segments and structures [3]. G. Tsontzos et. al. emphasized the feelings that help us to understand one other better, and it's only natural that this understanding should be extended to computers [4]. According to Mehmet Berkehan Akçay et. al., the objective of the neural networks is mostly utilized in industrial control and robotics applications [5]. However, the recent advancement in this technology has spread across various fields with higher accuracy.

According to Y. Wu et.al. discriminative testing has been utilized for voice detection along with the recognition for a longer period [6]. According to Varghese et al., there are numerous techniques to deduce feelings from formal as well as an informal expressions [7]. Many attempts have been made to use voice information to identify states. Some fundamental voice function vectors, in which utterance level data are measured, have been chosen to interpret feelings. A

study of the evaluation participants gave during the experiment confirmed a unique automated stimulus selection strategy based on effective tags [8]. There were significant correlations between participant assessments and EEG frequencies.

The scales of arousal, valence, and liking were classified in a single trial using data collected from various modalities. The findings were found to be much superior to those obtained from other classifications. Finally, decision fusion of these outcomes resulted in a slight performance improvement, demonstrating that the modalities are at least partially complementary. The database has been made available to the public, and we hope that other academics will use it to test their approaches and algorithms [9].

According to Peng et al., speaker identification refers to recognizing persons based on their speech. Because of its ease of use and lack of involvement, this technology is gradually being accepted and used as a form of biometrics, and it has quickly become a research hotspot in the field of biometrics [10].

The out-turned emotion labels are fused using majority voting to get the final emotion label for the given input voice signal. Furthermore, SVM and its combination approaches have been used in a lot of research. However, in the context of ensemble learning, the state can be mixed with other detection models. Furthermore, the best model for SER has yet to be determined.

III. ARCHITECTURAL BACKGROUND

The speech emotion recognition model was created utilizing deep learning techniques and machine learning methods. We will utilize the Python language because of its relatively large and distinct benefits. Some of the important libraries which are used in our simulation process are explained below.

A. Pandas Library

Panda is known as Panel data. It is a data manipulation and analysis software library created exclusively for this purpose. It gives us the data structures and operations we need to manipulate and construct the time series analysis along with the numerical tables. It is a complete open-source package.

B. Librosa Library

This is a python module for investigating the music and sound which gives us the referral devices we want to design and develop. It absorbs the sound documents as the input for the examination and modifies them with the help of Fourier transform techniques. It also plots the outcomes of the model.

C. Soundfile Library

In this python sound library various other libraries like NumPy, CFFI, etc. İntegrates themselves to utilize and compose the sound documents for altering different sound records in various formats.

Various representation models by the speech recognition system are mentioned as follows.

A. MFCC

The mel-recurrence cepstrum (MFC) is a representational module. It shows the input sound's short-term power spectrum based on the linear cosine transform. Mel Frequency Cepstral Coefficients (MFCC) are utilized to recuperate sound from a wav sound document utilizing different jump lengths and HTK-style Mel frequencies outline of the above-made sense of steps is plotted in figure 1 underneath.

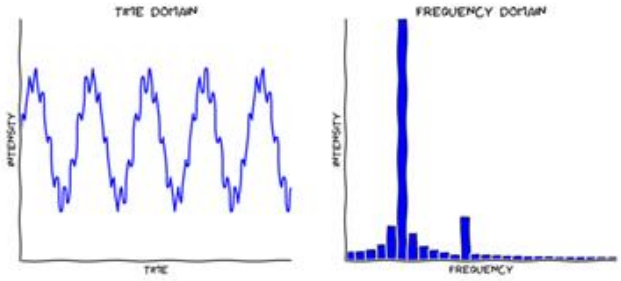

Figure. 1. Plot of Mel-frequency cepstrum

B. Mel Scale

As a general rule, a human being can hear the sound of 20Hz to 20kHz frequency. A tune of 300 Hz will be as similar as the sound of a standard dialer tone of a land-line telephone. Essentially, a 400 Hz will sound a piece higher. On contrasting the distance between these two howsoever this might be seen by your mind will be treated as the practically same [11]. Presently again the apparent distance of a 900 Hz signal and a 1kHz sound appears to be more prominent than the initial two albeit the real distinction is something very similar (100Hz). The Mel scale attempts to catch such contrasts. The fundamental recipe to change over the recurrence estimated in Hertz (f) into the Mel scale is given by:

$$\mathrm{Mel}(f) = 2595 \log\left(1 + \frac{f}{700}\right)$$

By the shape of a vocal track, the sound is generated by a human being. If the shape of the vocal tract can be resolved precisely then the addressing of the voice tunes or pitches will be more accurate. Embedding the time power range of the speech signal and MFCC can detect the voice. The detailed flowchart is displayed in figure 2.

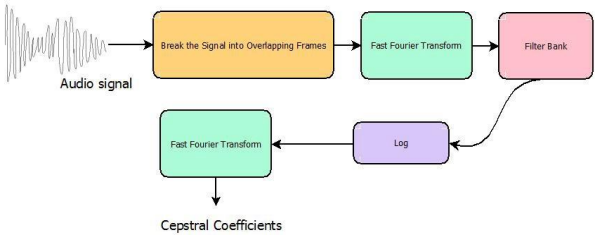

Figure. 2. Flowchart of MFCC

C. Spectrogram

A spectrogram is an integration of multiple FFTs stacked on top of one another [12]. It is a visualization approach of representing the loudness or the strength of the signals over a certain period concerning various frequencies present in a waveform.

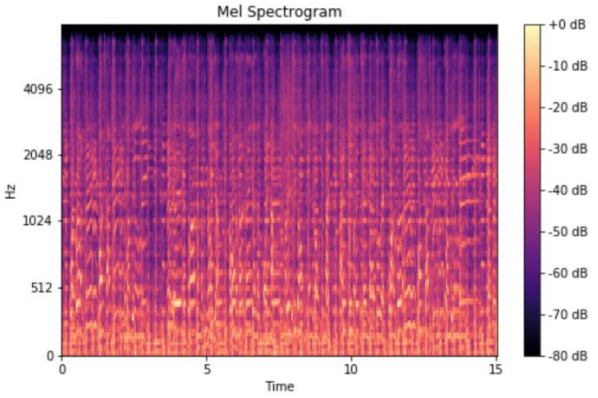

Figure. 3. Frequency vs Time diagram of Mel spectrogram

In this article, we have chosen the RAVDESS (Ryerson Audio-Visual Database of Emotional Speech and Song) dataset from Kaggle, which is a big dataset of audio recordings of talks in various moods, most of which were voiced by actors and actresses [13]. This dataset is 21 GB in size, but we compressed it to 30 MB. It has 7356 files, with 60 tries for each performer that contributes his voice. Speech files account for 2880, while Song files account for 2024 etc.

IV. METHODOLOGY

The major objective of this work is to create an elementary, efficient, and practical model that uses machine learning techniques as its basic feature, for which we can able to trust the system in producing accurate and error-free results. With the help of a python programming framework, named Librosa, designing and executing this work is pretty simple. The operation is heavily reliant on this language. Aside from writing the code, there are a few other parts to this work that are explored in depth. The technique is straightforward, and we've taken a conscientious approach to this work. This includes the addition of some of the new functionalities and features to the existing process of execution. The detailed flowchart is depicted in Figure 4.

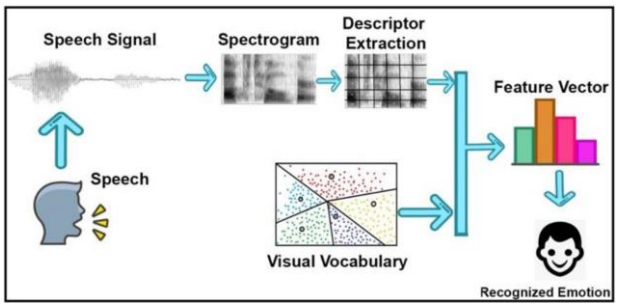

Figure. 4. Flowchart of the proposed model

This paper has two outcomes: the preliminary and the final. The first Outcome is nothing but the response that can be obtained from the identification of any sentiment in existence to the input fed to it. We created the system to access this response and do the additional trials by accumulating more samples and analyzing them to generate a pattern-based response. The system will be built based on this pattern. We will give random inputs to this system to test its functionality, response production, faults, etc. to discover the efficiency of the model.

We will assume that the system is performing considerably by the requirements if it creates appropriate answers for the provided loads. The efficiency percentage is used to assess how well things are working. On this basis, every system's efficiency, functionality, and dependability can be classified. The efficiency percentage of a general efficient system will range from 75 to 99 percent. We can deduce two things if it reaches 100 percent. As a result of the system's inability to manage the inputs, efficiency has been lowered to an unacceptably low level. Secondly, the model is either ineffective or inapposite as it is being analyzed by the human response.

A multilayer perceptron is made up of many perceptrons. As a result, the term "multi-layer" was coined. Regular artificial neural networks are similar to these multi-layer perceptron classifiers. A basic input layer in a multi-layer perceptron accepts the signal and an output layer that predicts the received input with the help of one or a random number of hidden layers in between these input and output layers which can able to forecast any steady function.

These classifiers can grasp and learn how to execute the work based on the data provided for training in the input layer. This is the most desirable strategy for recognizing speech patterns. The input variables for this classifier are not restricted in any way.

V. WORKING MODEL

To begin with, we must choose an environment in which to demonstrate the execution code. We have several possibilities for this, namely:

- Python IDLE
- Jupyter Notebook
- Anaconda Navigator
- Google Colab
- Pycharm

In this paper, we will look at Pycharm. It is a Python-designed environment for creating new code, editing existing code, and more. Following the selection of the environment, we will design and develop the execution code in this phase. One input data set is required in this stage. In this proposed work, the Ryerson Audio-Visual database of emotional speech and song (RAVDESS) dataset is chosen to be worked out. This dataset contains a total of 7356 files with 10 times on emotional validity and genuineness. The whole dataset is a size of 24.8GB from 24 actors. At first, it is downloaded from the Kaggle repository and saved in a local system as a principal area where it can be analyzed and the versions will be saved. This is for the system's simple access to the dataset during compilation.

Then the required libraries are being imported for fast execution in the code file. We should import them into the program even though they are installed on our PC. After saving it in our local system we are now able to gain access to the audio files which are being gathered for examination earlier. In the very next step by creating functions with various arguments like chroma, filename, mfcc, and mel the dataset can be read. It is very difficult to examine such a largescale dataset due to low processing capacity. As a result, we will use the looping concept to read each of the files. The program uses a variable called X to read all of the data in float32 format.

Based on the values obtained at the analysis on the chroma and The MFCC for each of the signals and if loop statement will be used to design certain patterns from each of the signals. The resultant of the If loop will be containerized in an array for further simulation.

VI. ISSUES AND CHALLENGES

Even though speech recognition frameworks have made considerable progress, there are as yet a couple of barricades in defeating to accomplish great acknowledgment. The creation of the dataset used in the learning system is quite possibly the main issue. Most of the SER informational collections are acted or evoked and recorded in assigned quiet rooms. Real-world data is boisterous and has fundamentally more particular properties than the other available data. Regular informational collections are additionally accessible, but there are fewer of them. The recording and use of regular feelings raise legitimate and moral issues nowadays. Most of the expressions in normal informational indexes come from syndicated programs, contact focus accounts, and different circumstances when the gatherings included know about the recording. This collection of information may exclude all the attached feelings and may not be reflected accurately. There are also issues with the marking of the expressions. So it is very challenging to detect the speaker's genuine inclination towards his or her speech.

Almost 90% identification can be done by the human annotator but not beyond that. For people, nonetheless, we feel that while deciding a speech, we should likewise think about the substance and setting of the correspondence atmosphere. SER is additionally affected by social and phonetic variables. There are different methods to examine cross-language SER that are accessible. However, the results uncover that the current frameworks and highlights are inadequate. For instance, the sound of feelings on speech in various dialects might vary. Whenever there are a few signal flags, the SER framework should figure out which sign to zero in on, and in which the issue is disregarded. Nevertheless, the way it tends to be overseen at the pre-processing stage with a speech partition algorithm, the current frameworks neglect to distinguish these issues.

VII. CONCLUSION AND FUTURE SCOPE

The whole system demonstrates the usages of machine learning to extract the ultimate emotions from audio data, as well as some insights into human emotion representation via voice. Thus, this technology can be used in multiple settings, including call centers for customer service in marketing, linguistic research, voice-based virtual assistants, etc. The following scopes are some of the procedures that can be taken to ensure that the models are well-formed and accurate:

- An exact implementation of the speaking speed can be investigated to see whether there is any inaccuracy, thereby resolving some of the model's flaws.
- Trying to figure out how to remove the audio clip's aimless stillness.
- Exploration of various acoustic features of sound data is being investigated to see if they may be used in the realm of speech emotion identification.
- Using an ensemble method and a lexical features-based approach to SER.

REFERENCES

[1] Puri, Tanvi, et al. "Detection of Emotion of Speech for RAVDESS Audio Using Hybrid Convolution Neural Network." Journal of Healthcare Engineering 2022 (2022).

[2] Xu, Mingke, Fan Zhang, and Wei Zhang. "Head fusion: improving the accuracy and robustness of speech emotion recognition on the IEMOCAP and RAVDESS dataset." IEEE Access 9 (2021): Pp. 74539-74549.

[3] Franciscatto, Maria Helena, et al. "Towards a speech therapy support system based on phonological processes early detection." Computer speech & language 65 (2021): 101130.

[4] Tsontzos, Georgios, et al. "Estimation of general identifiable linear dynamic models with an application in speech recognition." 2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP'07. Vol. 4. IEEE, 2007.

[5] Akçay, Mehmet Berkehan, and Kaya Oğuz. "Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers." Speech Communication 116 (2020): Pp. 56-76.

[6] Wu, Yonghui, et al. "Google's neural machine translation system: Bridging the gap between human and machine translation." arXiv preprint arXiv:1609.08144 (2016).

[7] Varghese, Ashwini Ann, Jacob P. Cherian, and Jubilant J. Kizhakkethottam. "Overview on emotion recognition system." 2015 International Conference on Soft-Computing and Networks Security (ICSNS). IEEE, 2015.

[8] Abbaschian, Babak Joze, Daniel Sierra-Sosa, and Adel Elmaghraby. "Deep learning techniques for speech emotion recognition, from databases to models." Sensors 21.4 (2021): 1249.

[9] Issa, Dias, M. Fatih Demirci, and Adnan Yazici. "Speech emotion recognition with deep convolutional neural networks." Biomedical Signal Processing and Control 59 (2020): 101894.

[10] Peng, Shuping, et al. "Remote speaker recognition based on the enhanced LDV-captured speech." Applied Acoustics 143 (2019): 165-170.

[11] Mao, Shuiyang, et al. "Revisiting hidden Markov models for speech emotion recognition." ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP). IEEE, 2019.

[12] Tarunika, K., R. B. Pradeeba, and P. Aruna. "Applying machine learning techniques for speech emotion recognition." 2018 9th International Conference on Computing, Communication and Networking Technologies (ICCCNT). IEEE, 2018.

[13] Wen, Guihua, et al. "Random deep belief networks for recognizing emotions from speech signals." Computational intelligence and neuroscience 2017 (2017).

[14] El Ayadi, Moataz, Mohamed S. Kamel, and Fakhri Karray. "Survey on speech emotion recognition: Features, classification schemes, and databases." Pattern recognition 44.3 (2011): 572-587.