HEAR-IT - Uncovering Emotions

Sathish G C

Harish Naidu Gaddam

Likhitha S

Thejesh Manduru

Manohar Kommi

Department of CSE

Bangalore Karnataka, India

Sathish.gc@reva.edu.in

harish.naidu0207@gmail.com

likhitha.srirangam@gmail.com

thejesh.m2000@gmail.com

kommimanoharnaidu@gmail.com

Abstract: SPEECH EMOTION RECOGNITION (SER) is that the act of trying to acknowledge human feeling an therefore the associated emotive states from speech. This takes advantage of the actual fact that tone and eat the voice of times mirror underlying feeling. In recent years, feeling recognition has been a chop-chop growing analysis domain. Machines, not like humans, lack the power to understand and specific emotions. However, by implementing automatic feeling recognition, human-computer interaction may be improved, reducing the necessity for human intervention. During this project, basic emotions like calm, happiness, fear, disgust, so on area unit extracted from emotional speech signals. We have a tendency to use machine learning techniques like the Multilayer Perceptron Classifier (MLP Classifier) that is employed to classify the given information into nonlinearly separated teams. The MLP classifier is trained mistreatment MEL, MFCC, CHROMA, and MEL options extracted from speech signals. To accomplish this goal, we have a tendency to use Python libraries like Librosa, sklearn, pyaudio, numpy, and soundfile to analyse speech modulations and acknowledge feeling.

Keywords*:* SPEECH EMOTION RECOGNITION, *MLP, MFCC, CHROMA, MEL, Librosa, Pyaudio, RAVDESS, SAVEE.*

---

1. INTRODUCTION

AI for emotion detection and analysis, often known as "affective computing," is a branch of artificial intelligence that focuses on human emotion recognition and analysis. Machines with this level of emotional intelligence are capable of grasping not just the cognitive channels of human communication, but the emotional channels as well. This provides children with the capacity to perceive, evaluate, and respond correctly to both verbal and nonverbal cues in a variety of situations. Researchers are putting in significant effort to teach robots to identify and understand human emotions, with which the field has made significant progress. Machine learning and deep learning are two technologies that are particularly important in this situation. In combination with these technical breakthroughs, images and speech recognition systems are utilised as inputs for the machines, which are then processed by the machines. Consequently, the robots learn to detect and interpret a grin or shift in tone of voice, such as whether it is a joyful or sad smile, for

instance. It has an influence on whether or not the current condition is better or worse than it was in the prior scenario. According to the researchers, characteristics such as skin temperature and heart rate are also being experimented with at this time. They are useful in the development of wearable devices that are as intelligent as possible, among other things.

2. RELATED WORK

[1] [2] They suggested a method that uses Random Deep Belief Networks to figure out what people are feeling based on what they say, which encapsulates the ensemble learning strategy used in the RDBN method for distinguishing emotions from voice data. The random subspaces technique was used once the input voice stream had been stripped down to its bare essentials. In this case, each random subgroup is supplied into the DBN input to capture the greater characteristics of the i/p file, which is after fed into the basic algorithm to obtain a projected mood classification. Additionally, each emotion label output is fused by a clear majority to create the final emotion tag for the given audio signal.

[3] [4] [5] They suggested the use of a feeling detection mechanism in conjunction with a deep learning approach. Despite the fact that large amounts of acoustic emotional data are used to describe emotions.

[9] suggested an emotion identification technique based on neural networks that focuses on distinguish emotions from a given i/p audio stream using NN Algorithms. They recommended an elevated pass filter as the first step in this strategy.

[10] In this article, the author projected a method for recognising negative emotions in the Thai language using deep learning. They used two 2-D convolutional neural networks (CNNs) and trained their models individually using the RAVDESS, TESS, SAVEE, and Crema-d datasets. Finally, they evaluated their model using the Thai dataset.

[11] In this author's project, using attention head fusion is used to increase the accuracy of emotion identification in speech. They used the head fusion approach to create a feature with MFCCs as input characteristics to construct an ACNN algorithm for emotion recognition in speech. They confirmed their findings by examining the IEMOCAP database, which includes information on 4 emotions (angry, sad, excited, and neutral).

[12] projected a method for identifying emotions in speech called "Direct Modeling of Speech Emotion from Raw Speech." They used a mixture of CNN and LSTM. They extracted characteristics from raw speech using parallel convolutional layers with variable filter lengths.

[13] proposed speech emotion identification starting with a log Mel spectrogram, the researchers utilised the Librosa package to extract features from it. These characteristics were then used to create textures on the time and frequency axes by utilising two concurrent convolutional layers. As a result, an 80-channel model is divided into 4 convolutional layers, which would then be provided in the fifth layer. As a consequence, the attention layer responds to the representation and conveys the findings to the fully connected layer, which is responsible for ultimate emotion categorization. In addition, a great deal of research has been done utilising SVM and its combination approaches to date.

[14] However, for voice emotion classification, we combined random subdomains, MLP, and CNN to build an ensemble teaching model. Additionally, the optimal model for SER has not been investigated.

3. METHODOLOGY

In this case, we will partition our database into 2 parts: training and testing. After making a partition of our database, we'll load it and execute 2 processes: first, we'll extract the dataset's features, and then we'll use a range of classifiers to determine the particular emotion reflected by the i/p audio file. After training and testing our model, then feature extraction and then implementing the classifier on the data is integral to maintaining that the classifier is as precise as possible. Whichever classifier has the best accuracy, we need to save it in order to implement it using the Flask framework. The user may deliver the audio file using this web application. This web application will connect to the recommended model and use the given audio sample to determine the exact emotion. It will then play music in accordance with the emotion found in an attempt to lift the user's mood.

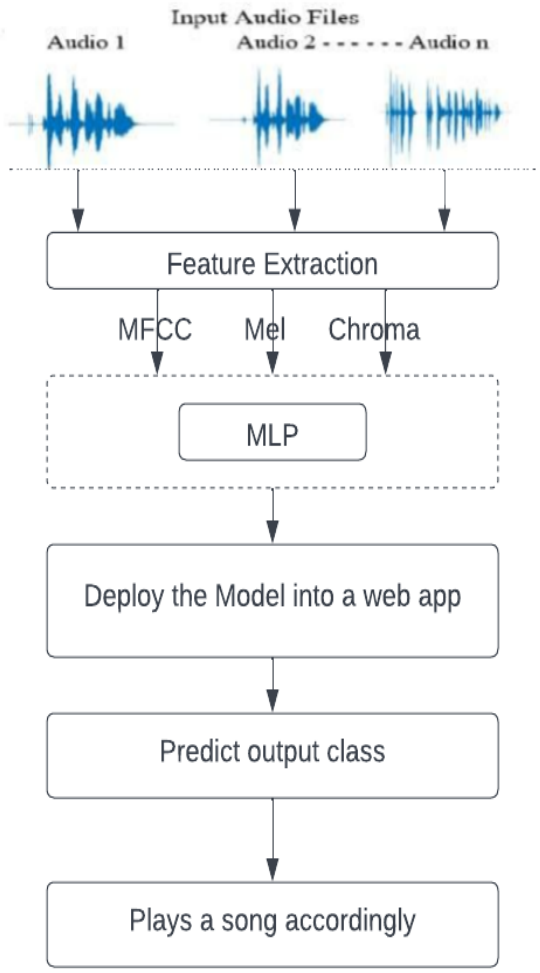

Fig 1 : System Architecture Diagram

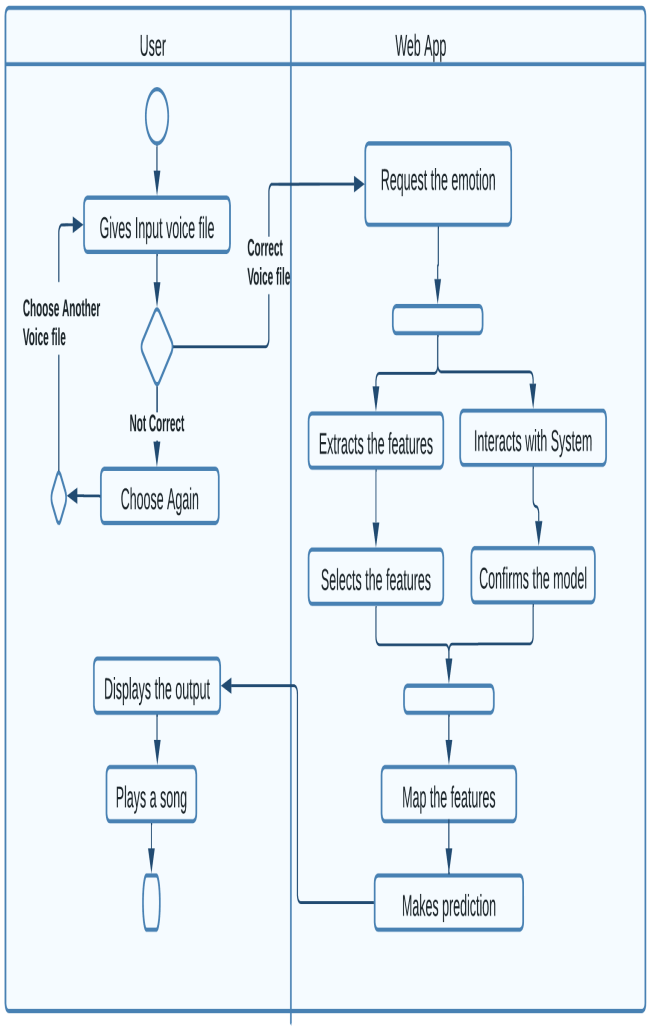

Fig.2 Activity Diagram

In the fig 2 indicates that the user first submits an audio clip file through the web application. By clicking on the Predict button and providing the appropriate voice input file, the user may request an emotion from the online server model. If the candidate i/p file is erroneous or if the user selects an unsupported voice file, the user must choose another supported and suitable voice file. After submitting the appropriate input file, the user may query the web application for the emotion associated with that particular voice input file. The web application communicates with the system and saves the given input voice file, takes features from it, and retrieves the stored model. The web application maps the selected data to a certain emotion using the previously stored model and then returns the information for the given input audio clip that was entered into the system.

2. Extraction of Characteristics

In this case, transforming our supplied i/p audio recordings to programmed data as the models will be unable to grasp the data. To evaluate these artefacts, we leverage the librosa package, which has different packages for collecting characteristics from an acoustic file's data. We investigate the MFCC characteristic since it is the most useful and vital approach to do so. It does so by examining the transmitter's short-term power spectra rather than its long-term prospective. The second feature considered is the Mel feature. The 3rd feature that is looked at is the Chroma feature.

The ZCR characteristic is used to figure out how quickly a certain frequency changes sign over the course of a frame, and the Rms characteristic is used to figure out how loud the wav file is.

3. Selection of Feature

Every one of these qualities are assessed in respect to the source audio clip's frequency, pitch, and intensity. We need not analyze every aspect of the MFCC to get satisfactory results. To do this, we evaluate Mfcc's top tier characteristics; then evaluate 12 Chroma characteristics, 128 Mel characteristics, 1 Zcr characteristic, and 1 Rms characteristic. A range of 182 characteristics are being evaluated, which includes identifying and picking 182 characteristics from every file, including them all to the x list, adding all of the emotions to the y list, and transforming the x list to array called q, providing it to the model through train test partition. We are also examining the following characteristics: Where p is the impartial characteristic, and y represents the heavily reliant characteristic, which encompasses emotions.

4. Developing a web-based Application

We created a web application using the Flask framework. Flask is free and open-source software that makes applications run on a server. Flask includes a templates folder, which is a static directory in which we may keep the HTML, CSS, Java Script, and image files for our web application.

4. SYSTEM REQUIREMENTS

TABLE I

Software	Hardware
OS: Windows 7 ABOVE. Coding Language: Python.	System: Intel i3 2.4 GHz and above.
IDE: Visual Studio Code	Hard Disk: 100 GB.
Front End: HTML/CSS, Flask.	Monitor: 15 VGA Colour RAM: 4 GB.

5. OBJECTIVES

- The primary goal of this project is to improve the interaction between humans and machines.
- Building a website with an excellent user interface and functionality.
- Deliver clean and accurate Emotion detection, Plays a music that is associated with improving the mood of the individual whose emotion has been recognized.

6. APPLICATIONS

- Interaction between humans and computers.
- Autonomous vehicles and smart home automation.
- Medical – Psychiatrists, Autism Spectrum Disorder.
- Hear It software takes advantage of music's inherent mood-lifting properties to aid individuals in improving their mental health and overall well-being.

7. CONCLUSION

By developing this project, we can utilise machine learning to recognise the emotion in speech, which may then be used to improve human-computer interaction. This method may be used to improve virtual voice-based assistants who can comprehend human emotions and respond accordingly, as well as in marketing and enhancing customer service in contact centers. With this approach, we acquire an approximate accuracy of roughly 80%. We utilised the MLP classifier approach for recognising emotions in this research, as well as speech as input. We discovered that MLP models worked better than others, and we used the Flask framework to deploy that MLP model into a web application.

8. REFERENCES

[1] Abhishek, Kalyani, Vaishnav Sham 2021 Emotion Based Music Player, International Journal of Computer Science and Mobile Computing, Vol.10 Issue.2, February- 2021, pg. 50-53.

[2] Rajdeep Chatterjee, Saptarshi Mazumdar, R. Simon Sherratt Real-Time Speech Emotion Analysis for Smart Home Assistants, IEEE TRANSACTIONS ON CONSUMER ELECTRONICS, VOL. 67, NO. 1, FEBRUARY 2021.

[3] Advait Gopal Ranade, Maitri Patel, Archana Magare 2018 Emotion Model for Artificial Intelligence and their Applications, 5th IEEE International Conference on Parallel, Distributed and Grid Computing(PDGC-2018), 20-22 Dec, 2018, Solan, I.

[4] M. Aravind Rohan, K.Sonali Swaroop, B. Mounika K. Renuka, S.Nivas. 2020 EMOTION RECOGNITION THROUGH SPEECH SIGNAL USING PYTHON ,978-1-7281-7213-2/20/$31.00 c©2020 IEEE

[5] D. Bharti and P. Kukana, “A Hybrid Machine Learning Model for Emotion Recognition From Speech Signals”, In 2020 International Conference on Smart Electronics and Communication (ICOSEC).IEEE, pp. 491-496, September 2020.

[6] Mingke Xu et al. “Speech Emotion Recognition with Multiscale Area Attention and Data Augmentation”, ArXiv abs/2102.01813,February 2021.

[7] Abhay Kumar et al. “Speech Mel Frequency Cepstral Coefficient feature classification using multi level support vector machine”, In 2017 4th IEEE Uttar Pradesh Section International Conference on Electrical, Computer and Electronics (UPCON). IEEE, pp. 134-138, October 2017.

[8] G. Deshmukh, A. Gaonkar, G. Golwalkar and S. Kulkarni, “Speech based Emotion Recognition using Machine Learning”, In 2019 3rd International Conference on Computing Methodologies and Communication (ICCMC). IEEE, pp. 812-817, March 2019.

[9] Z. Tariq, S. K. Shah and Y. Lee, “Speech Emotion Detection using IoT based Deep Learning for Health Care”, In 2019 IEEE International Conference on Big Data (Big Data). IEEE, pp. 4191- 4196, December 2019.

[10] Livingstone SR, Russo FA, The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodalset of facial and vocal expressions in North American English. PLoSONE 13(5): e0196391, May 2018.

[11] Manas Jain et al. “Speech Emotion Recognition using Support Vector Machine”, arXiv:2002.07590, February 2020.

[12] Mandeep Singh, Yuan Fang, “Emotion Recognition in Audio and Video Using Deep Neural Networks”, arXiv:2006.08129, June 2020.

[13] P. Shen, Z. Changjun and X. Chen, “Automatic Speech Emotion Recognition using Support Vector Machine”, In Proceedings of 2011 International Conference on Electronic & Mechanical Engineering and Information Technology. IEEE, pp. 621-625, August 2011.

[14] K. Tarunika, R. B. Pradeeba and P. Aruna, “Applying Machine Learning Techniques for Speech Emotion Recognition”, In 2018 9th International Conference on Computing, Communication and Networking Technologies (ICCCNT). IEEE, pp. 1-5, July 2018.