Category: Technical Paper

HEAR-IT - Uncovering Emotions

Sathish G C
Department of Computer Science & Engineering
Bangalore Karnataka, India
Sathish.gc@reva.edu.in

Gaddam Harish Naidu
Department of Computer Science & Engineering
Bangalore Karnataka, India
R18CS134@cit.reva.edu.in

Likhitha S
Department of Computer Science & Engineering
Bangalore Karnataka, India
R18CS203@cit.reva.edu.in

Manduru Thejesh
Department of Computer Science & Engineering
Bangalore Karnataka, India
R18CS214@cit.reva.edu.in

Kommi Manohar
Department of Computer Science & Engineering
Bangalore Karnataka, India
R18CS184@cit.reva.edu.in

Abstract: Speech feeling Recognition (SER) is that the act of trying to acknowledge human feeling and 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-frequency cepstrum coefficients (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: SER, MLP Classifier, MFCC, Chroma, Mel, Librosa, Pyaudio, RAVDESS, SAVEE., Python, Flask, HTML, CSS, JavaScript.

I. Introduction

AI for emotion detection and analysis, often known as affective computing, is a subfield of artificial intelligence that is concerned with the detection and analysis of human emotions. 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, which is an important step forward in the area. 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 utilized as input 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.

II. Related work

Guihua Wen et al. [1][2] suggested a Random Deep Belief Networks for Recognizing Emotions from Speech Data, which covers the ensemble learning approach of the Random Deep Belief Networks (RDBN) method for recognising emotions from speech signals. Where they first retrieved the low-level properties of the supplied input voice stream so used the Random subspaces approach. Where each Random subspace is fed into the input of DBN to extract the higher-level characteristics of the given input speech signal and delivered as input to the bottom classifier to supply a projected emotion label. Furthermore, each outputted emotion label is fused by majority voting to work out the ultimate emotion label of the given input speech signal.

M. Shamim Hossain and Ghulam Muhammad [3][4][5] advised an feeling detection system supported a deep learning technique victimisation Audio-Visual emotional massive information, describing however emotions could even be recognised victimisation voice and video as input. [6] employed two datasets for this purpose: one may be a Big Data database comprising both speech and video input files, and therefore the other is that the eNTERFACE database. During this method[7][8,] they first extracted the characteristics of the supplied input voice signals to form a Mel-spectrogram, which can be regarded an image. This Mel-spectrogram is put into a 2D CNN, which is then fed into extreme learning machines (ELMs) for score fusion. within the case of video signals, certain sample frames from a video segment are retrieved and fed into the 3D CNN, which is then followed by extreme learning machines (ELMs) for score fusion. The output of both of those speech and video fusions is fed into an SVM for final emotion categorization of the provided input speech and video signals.

Pawan Kumar Mishra and Arti Rawat [9] proposed a “ Speech emotion detector, which mainly aims on a way to differentiate emotions from a given input audio signal using Neural Networks”. during this method, they first proposed a High Pass Filter, which is an electronic filter wont to remove unwanted sounds from a given input speech signal, i.e., which passes only the frequency that's higher and when there's a coffee frequency interrupt frequency occurs and which passes only the frequency that's higher for further performing feature extraction process. They employed MFCC to extract features. Following feature extraction, they employed a neural network to categorise the ultimate emotions from the supplied input audio data.

Sakorn Mekruksavanich, Anuchit Jitpattanakul, and Narit Hnoohom [10] suggested a “Emotion recognizer for Thai language using neural networks, where they used two one-dimensional Convolutional Neural Network (CNN), they trained their model individually using RAVDESS, TESS, SAVEE, Crema-d datasets for classifying whether the emotion is positive or negative”. Finally, they ran their model through its paces on the Thai dataset.

Mingke Xu et al. [11] proposed using Attention Head Fusion to boost the accuracy of speech emotion recognition, where they implemented an ACNN model using head fusion technique to come up with a feature point Xfusion, using MFCCs as input features to recognise emotions in speech and validated their result using IEMOCAP dataset, considering four emotions (angry, sad, excited, and neutral) for emotion recognition. Siddique Latif et al. [12] suggested a "Direct Modeling of Speech Emotion from Raw Speech," within which they employed a mixture of Convolution Neural Networks (CNNs) and Long Short Term Memory (LSTM) for emotion identification. during this study, they initially employed parallel convolutional layers with varying filter lengths to extract features from raw speech, and so concatenated all of those parallel convolution layers with a 2D convolution layer using max-pooling. Mingke Xu et al. [13] suggested speech emotion recognition using multiscale area attention and data augmentation, during which they used multiscale area attention to SER and created an attention-based convolutional neural network. First, they utilised the Librosa package to extract the log Mel spectrogram as features, which were then input into two concurrent convolutional layers to make textures for the time and frequency axes, respectively. The result's an 80-channel representation that's fed into four successive convolutional layers. Following that, the eye layer responds to the representation and passes the findings to the fully connected layer for ultimate emotion categorization. Furthermore, a considerable amount of research has been conducted utilising SVM and its combination methods [15][16][17]. Perhaps, the current trend in technology does not have the combination of random subspace, MLP, and CNN for speech emotion identification in a learning framework. Furthermore, the optimum model for SER has not yet been examined.

III. DATASET DESCRIPTION

Throughout the course of our study, we relied on three different datasets. One of them is the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) dataset[18], which contains 1440 audio-speech files belonging to 24 professional actors, 12 of whom are male and 12 of whom are female, with 60 trails for each of them totaling 1440 audio-speech files, and 1012 audio-song files belonging to 24 professional actors, 12 of whom are male and 12 of whom are female, with 44 trails for each of them totaling 1012 Whereas Audiospeech files have sad, calm, angry, happy, afraid, disgusted, and surprised expressions, Audio-song files include sad, calm, angry, fearful, and joyful emotions.

The second dataset we looked at was the Toronto emotional speech set (TESS) dataset, which contains 2800 unique voice files recorded by two actresses aged 26 and 64 years, each of whom say a set of 200 target words in the carrier phrase "Say the word ." Both women are from Toronto, speak English as their first language, have a university education, and have musical training. There are approximately 1401 unique speech files for those aged 26 to 64, and around 1399 unique speech files for people aged 65 to 74. When we combine both of these voice files, we get 2800 distinct speech files with seven different emotions for each of them: anger, contempt, fear, pleasure, pleasant surprise, sorrow, and neutral.

The final dataset we looked at was the Surrey Audio-Visual Expressed Emotion (SAVEE) dataset [[20], which contains seven different emotions recorded by four native English male speakers, DC, JE, JK, and KL. They range in age from 27 to 31 years old and are postgraduate students and researchers at the University of Surrey. This dataset comprises 480 distinct audio recordings containing seven different emotions: anger, contempt, neutral, fear, happiness, sorrow, and surprise. We have around 5732 distinct voice recordings after integrating all three datasets. By merging these three datasets, we were able to observe eight different emotions for our study. They are peaceful, neutral, pleased, sad, furious, afraid, disgusted, and shocked.

IV. PROPOSED METHODOLOGY

The first stage will be to break down our dataset into two steps: training and testing our model. As soon as we've divided our dataset, we'll need to load it and run two processes: first, we'll extract the features from the dataset, and then we'll use a variety of classifiers to determine the precise emotion emitted by the given input speech signal. It is necessary to assess the correctness of our model once we have finished training and testing it by doing feature extraction and then applying the classifier to the data. Whichever classifier has the highest accuracy, we need to store that particular model in order to be able to deploy that model into a WebApp using the Flask framework. The audio file can be provided by the user through the usage of this WebApp. This web application will interact with the suggested model and detect the precise emotion from the provided input audio file. It will then play music in accordance with the emotion discovered in order to improve the mood of the person using the online application.

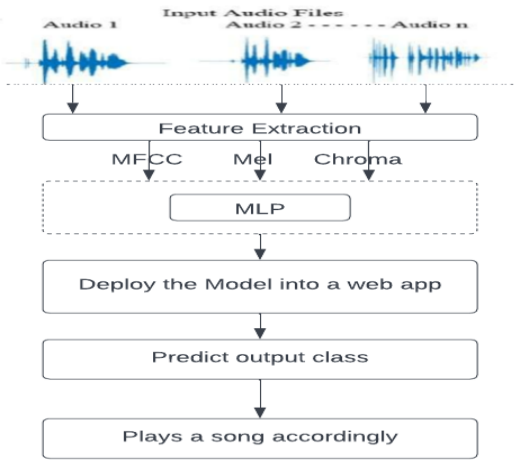
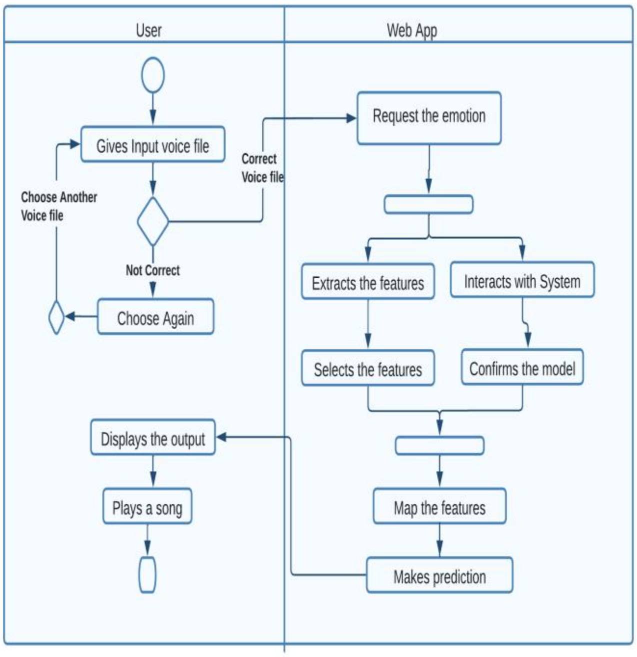


Fig: System Architecture Diagram

A. Activity Diagram



According to the activity diagram (Fig. 2), the user initially provides a voice input file via the web app. If the user provides the proper voice input file, he can request emotion from the web program by clicking the Make Prediction button. If the user's input file is incorrect or he or she has picked an unsupported voice file, the user must select another supported and appropriate voice file. After providing the relevant input file, the user can ask the web app for the emotion of that specific voice input file. By communicating with the system, the web app stores the supplied input voice file in the system, extracts features from it, and obtains the saved model path. Using these selected features and the saved model, the web app maps the selected features to a certain emotion by using the saved model to make a prediction, and then displays the outcome for the provided input voice file.

B. Feature Extraction

We are converting our given input audio files to digital data in Feature extraction because the models will not be able to understand the audio data, so we are transforming our given input audio file to digital data so that the model will be able to understand this digital data for this we are using librosa module which contains various libraries for extracting the features from the given input audio file by using sample rate and sample data. First, we consider the Mel Frequency Cepstral Coefficient (MFCC) feature, which is the most important and effective method for performing this feature extraction task. It analyses the signal based on short term power spectrum, i.e. by segmenting the speech sample into a number of frames by framing, then by applying certain window to reduce signal discontinuities at the beginning and end of each frame, and finally by using FFT to identify frequency spectrum of the signal. The second feature under consideration is the Mel feature, which will be utilized to capture the frequency characteristic of the supplied input audio file signal as expressed on the Mel scale. The third feature under consideration is the Chroma feature, which is used to capture melodic and harmonic characteristics of sound based on the pitch of the given input audio file, the Zcr feature, which is used to specify the rate of sign changes of the particular signal during the duration of the particular frame, and the Rms feature, which is used to analyze the loudness in the given input audio file because changes in loudness are important for extracting features in new input audio files.

C. Feature Selection

All of these characteristics are taken into account based on the energy, pitch, and loudness of the input audio file. We should not evaluate all of the aspects in the Mfcc in order to achieve decent outcomes. For this purpose, we evaluate just the first 40 features in Mfcc for filtering, which contribute the most to the classifier; we consider 12 Chroma features, 128 Mel features, 1 Zcr feature, and 1 Rms feature. We are extracting and selecting 182 features from each of the given input files, appending all of this to x list, appending all of the emotions to y list, and then converting x list into an array named p and passing it to the model via train test split. Where p represents the independent feature, which includes features, and y represents the dependent feature, which contains feelings.

D. Multilayer Perceptron (MLP) Classifier

All of these characteristics are taken into account based on the energy, pitch, and loudness of the input audio file. We should not evaluate all of the aspects in the Mfcc in order to achieve decent outcomes. For this purpose, we evaluate just the first 40 features in Mfcc for filtering, which contribute the most to the classifier; we consider 12 Chroma features, 128 Mel features, 1 Zcr feature, and 1 Rms feature. We are extracting and selecting 182 features from each of the given input files, appending all of this to x list, appending all of the emotions to y list, and then converting x list into an array named p and passing it to the model via train test split. Where p represents the independent feature, which includes features, and y represents the dependent feature, which contains feelings.

E. Creating a web API

We built a web application with the Flask framework. Flask is a web framework that provides tools, libraries, and technologies for creating an online application. Flask gives us with a templates folder, which is a static folder in which we may add the HTML and CSS files required by the web application. We can develop a web API by using the Flask framework's folders, tools, and libraries.

V. SYSTEM REQUIREMENTS

TABLE I

Software Requirements	**Hardware Requirements**
Operating System: Windows 8 ABOVE. Coding Language: Python, Machine Learning Algorithm.	System: Intel i3 2.4 GHz and above.
IDE: Visual Studio Code	Hard Disk: 100 GB.
Front End: HTML/CSS, Flask, JavaScript .	Monitor: 15 VGA Colour RAM: 8 GB.

VI. 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.

VII. 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.

VIII. CONCLUSION

By developing this project, we can utilize Machine Learning to recognize the emotion of the 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 utilized the MLP classifier approach for recognizing emotions in this research, as well as the speech as input. We discovered that MLP worked better than others, and we used the Flask framework to deploy that MLP model into a web app. We trained and tested our models on three datasets: the RAVDESS dataset, the TESS dataset, and the SAVEE dataset. Finally, we aggregated all three datasets into one in order to maximize the training data, and we assessed our model using this combined dataset. We can now integrate our deployed web app to any website, such as a customer support website or other websites where they wish to recognize their emotions and act or reply accordingly, because we have deployed our final MLP model into a web app.

In Future to improve real-time identification with customers, this work will be expanded to include more subjects, nations, accents, age and gender categories, and build an application that might be used in the field. We will also try to extract several auditory characteristics to utilize in sophisticated SER classification algorithms. Other variables will be introduced to assess a classification model's synthetic performance. We want to expand the training data in the future by adding more datasets like these three. However, we must ensure that the naming convention in the new datasets matches the three previously added datasets, and we can undertake certain augmentation procedures to improve the training data. We'd want to test various architectures on this merged dataset to increase model precision. However, we want to see if we can identify emotions utilizing video, picture, text, and speech inputs as well. This will help in applications where emotion detection is required to act or respond appropriately. In the future, we'd like to test numerous emotions from a single input file, as people's moods might change while speaking. Currently, our model can only recognize one emotion from a spoken sample.

IX. 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.

[15] Siddique Latif et al. "Direct Modelling of Speech Emotion from Raw Speech", Proc. Interspeech 2019, 3920-3924,September2019