Speech Emotion Recognition with Machine Learning

Katikireddy Lavanya
PG scholar, Department of MCA, CDNR collage, Bhimavaram, Andhra Pradesh.
A.Durga Devi
(Assistant Professor), Master of Computer Applications, DNR collage, Bhimavaram, Andhra Pradesh.

Abstract:

Understanding a person's emotion from their speech is called speech emotion recognition. It enhances interactivity between people and machines. Although it is tough to annotate audio since emotions are subjective, Speech Emotion Recognition (SER) makes it possible to forecast a person's emotional state. This is the same principle that dogs, elephants, horses, and other animals use to understand human emotion. There are several ways to gauge someone's emotional condition, including behaviour, expression, pitch, tone, etc. Some of these are supposed to enable the recognition of speech emotions. The classifiers are taught to recognise speech emotions using a limited amount of data points. This study considers the Ryerson Audio-Visual Database of Emotional Speech and Song dataset. Here, the three essential properties are retrieved, including chroma, Mel Spectrogram, and MFCC (Mel Frequency Cepstral Coefficients).

Keywords: Emotion, Machine learning, speech recognition.

I. INTRODUCTION

Speech is an essential emotional transporter in human communication. Speech-Emotion Recognition (SER) has numerous applications in robots, mobile services, psychological assessment, and other fields. A person's physical characteristics, such as tone of voice, breathing, heart rate, blood pressure, skin elasticity, muscle tension, etc., are influenced by their emotions.

The mimicry and expression of the face, the tone, and the pitch of the voice are only a few examples of how some of these physical reflections of emotions are much more obvious and externally accessible than others. The systems need to recognize the feelings in a speech to communicate effectively with people. Therefore, to have practical, clear communication like humans, it is necessary to create machines that can recognize paralinguistic information such as emotion.

Numerous machine learning algorithms have been created and tested to classify these emotions conveyed by speech. The goal of creating machines that can decipher non-linguistic information, such as emotion, aids in human-machine interaction and help to make the interaction more natural and clear. Convolutional neural networks are employed in this study to forecast the emotions in a speech sample.

The topic of research known as speech emotion recognition" (SER) focuses on creating methods and algorithms to recognize and categorize emotions expressed through voice signals automatically. SER has attracted a lot of attention due to developments in machine learning and signal processing because of its potential use in various industries, such as entertainment, customer service, healthcare, and human-computer interaction.

Since emotions are complicated and subjective phenomena impacted by different factors like intonation, pitch, speech tempo, and spectral features, understanding human emotions from the speech is complex. However, machine learning methods have demonstrated encouraging results in accurately training models to recognize various emotional states and extracting significant characteristics from speech data.

Speech and emotion recognition often involves several steps. First, noise is removed from voice data, volume levels are normalized, and essential features are extracted. Mel-frequency cepstral coefficients (MFCCs), prosodic qualities, and statistical characteristics can all be included in this list. These features record the voice signal's acoustic qualities, which can be used to classify emotions.

As part of the training process, the model parameters are optimized based on the labelled data provided to discover patterns and connections between the retrieved characteristics and corresponding emotions. By extracting the same set

of features and using them to classify the trained model, the model may be used to predict the sentiment of unseen speech signals.

Speech and emotion recognition have important ramifications across many industries. SER can improve user engagement and enjoyment by enabling computers to modify their replies based on the user's sensed emotional state during human-computer interaction. Regarding customer service, SER can help with sentiment analysis, allowing businesses to assess client happiness and swiftly rectify issues. In medicine, SER can aid in identifying and following up on emotional problems.

The heterogeneity in emotional expression across loud speech environments, cultures and people and the requirement for sizable annotated datasets for vital model training remain despite the advances made in SER. Current research focuses on overcoming these difficulties and investigating cutting-edge machine learning methods, such as multimodal strategies incorporating speech with additional modalities, including facial expressions and physiological information.

Automatic detection and classification of emotions expressed through speech is made excitingly possible by speech emotion recognition using machine learning. By enabling more sympathetic and individualized interactions between technology and people, SER has the potential to revolutionize several industries with additional breakthroughs in algorithms and data accessibility.

II. LITERATURE SURVEY

In recent years, emotion detection in speech has become an important topic of study. Reviewing previous research on the processing of emotional speech helps conduct additional studies in this area. The most recent research on speech emotion recognition is discussed in this work, considering the problems with emotional speech corpora, various speech features, and models for emotion recognition from speech. In this paper, 32 typical speech databases are evaluated in terms of their goal of collection, emotional range, speaker count, and language. Also briefly highlighted are the problems with emotional speech databases used in emotional speech recognition. The literature on various aspects applied to the challenge of voice emotion identification is given. The significance of selecting different classification models has been explored along with the review. Wherever relevant, the key points for future emotion identification research have been emphasized, generally and specifically for the Indian setting. [1]

The paper presents an experimental investigation into voice emotion recognition. As computer-generated figures such as avatars and virtual chat faces become increasingly prevalent, using emotion to guide their expression becomes more important. The study uses a corpus of emotional speech with 721 brief utterances that reflect the four emotions of neutrality, sadness, happiness, and anger. These utterances were manually extracted from movies and television plays. We present a brand-new idea to assess speech emotions. Since most spoken sentences cannot be accurately categorized according to emotion, most emotional states can nonetheless be defined as a combination of various emotions. We have created an agent that can recognize and express emotions and taught SVMs to recognize utterances within these four categories. [2]

Over the past two decades, research into automatic emotion recognition based on speech has expanded significantly. There are numerous approaches available for identifying emotional tones in speech. In this study, we take a look back at the multiple classifiers that have been used to recognize emotions in speech. Emotion classifiers categorize experiences as neutral, disgusted, afraid, sad, pleased, angry, surprising, etc. Emotional speech samples are employed as a database, and prosodic and spectral properties such as speech rate, formants, pitch, Energy, Mel frequency cepstrum coefficient (MFCC) and linear prediction cepstrum coefficient (LPCC) are extracted from these samples for use in emotion classification from speech. Classification accuracy is reflected by the features derived from the data. The benefits and effectiveness of a voice emotion recognition system that employs various classifiers are also examined. [3]

This study aims to label emotional states in spoken language as happy, sad, angry, or scared. The samples used in this study come from the Linguistic Data Consortium (LDC) and the University of Georgia database. Speaker rate, LPCC coefficients, MFCC coefficients, pitch, and Energy are all crucial parameters that can be calculated from the data. Support Vector Machine (SVM) is employed as the classifier to categorize these feelings, with Gender-Dependent Classification and One against All (OAA) being the two classification techniques. These algorithms, along with the MFCC and LPCC ones, have been compared and contrasted. [4]

Clear communications, which can be about anything, and implicit communications, which are indications about the speakers, make up the two channels via which people communicate. The first explicit channel has received much attention from linguistics and technology, but the second channel has received less attention. One of the primary responsibilities associated with the second implicit channel is comprehending the other person's feelings. To accomplish that goal, signal processing and analysis techniques must be developed, and linguistic and psychological evaluations of emotion must be combined. In those areas, this article discusses fundamental problems. [5]

The goal of emotional speech recognition is to automatically categorize speech units into emotional states such as sadness, neutrality, happiness, anger, and surprise. This paper's main contribution is an evaluation of the discriminating power of a set of characteristics for emotional voice recognition, taking into account the speaker's gender. Over 500 expressions from the Danish Emotional Speech Database were analyzed, and 87 factors were computed. The top 5–10 features for classifying samples according to gender have been identified using the sequential forward selection (SFS) technique. Bayes classifiers using Gaussian pdfs achieve 61.1% accuracy in classifying male subjects and 57.1% in classifying female subjects. A random classification method would only be 20% accurate in the same trial. A proper classification score of 50.6% is achieved when gender information is omitted. [6]

Six speech emotions— disgust, happiness, fear, surprise, anger, and sadness—are categorized from coarse to fine using a suggested three-level speech emotion identification model to address the issue of speaker-independent emotion recognition. Fisher rate, also considered an input parameter for support vector machines (SVM), is used to narrow down feature candidates from 288 at each level. Four comparative tests, comprising the Fisher's linear model (Fisher + SVM), artificial neural network (ANN) for classification, and principle component analysis (PCA) for dimension reduction, are designed to assess the suggested system. The experimental results showed that SVM performed better than ANN at recognizing the emotions of speeches delivered by anyone in the room. In contrast, Fisher performed better than PCA at dimension reduction. Average recognition rates across all three tiers are 86.5%, 68.5%, and 50.2%. [7]

III. PROPOSED METHOD

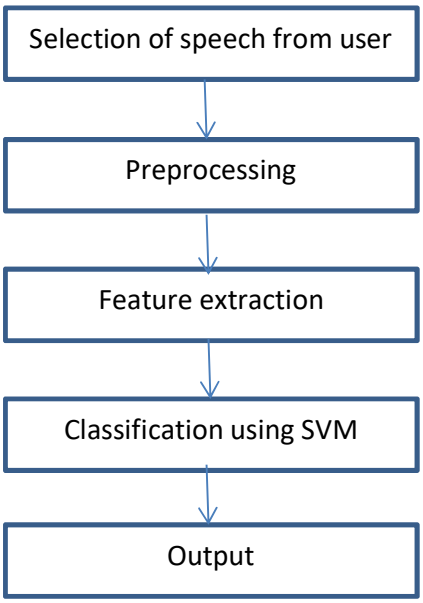

Fig. Block diagram

Following are the steps for implementing the system:

1. **Selection of speech from user**

First, select the speech form of the user that we want to recognise.

2. Pre-processing

Speech-emotion recognition (SER) with machine learning requires pre-processing. It entails converting and altering raw speech signals to extract useful features for use as input into machine learning models. The pre-processing strategies try to improve the input data's quality and discriminative strength, making it appropriate for emotion recognition tasks.

There are following steps for pre-processing

i. Sampling
ii. Pre-emphasis
iii. De-silencing
iv. Framing
v. Windowing

3. Feature extraction

Feature extraction is a fundamental step in machine learning-based Speech Emotion Recognition (SER). It entails choosing and extracting essential auditory or language elements from speech data that capture the underlying emotional content. These characteristics are fed into machine learning models for training and classification.

There are two methods to extract feature,

i. Energy Feature Extraction for each Frame
ii. MFCC feature vector extraction for each frame

4. Classification using SVM

For classification the Support Vector Machine linear classification algorithm is used. SVM is a nonprobability linear classifier. It is a classification algorithm that distinguishes between two classes. As a result, we create models that compare one feeling to another.

IV. RESULT

In this project we are detecting emotion using speech data and to implement this project we have trained CNN algorithm with RAVDESS Audio Dataset for speech emotion recognition. Below screen shots code with red colour comments showing extraction of MFCC features from audio dataset

To run project double click on 'run.bat' file to get below screen

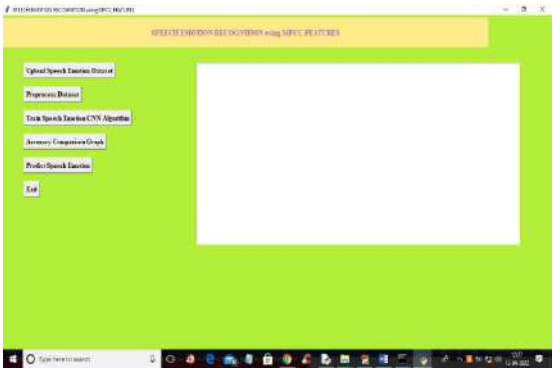

Fig.4.1 Run.bat file

In above screen click on 'Upload Speech Emotion Dataset' button to upload dataset to application

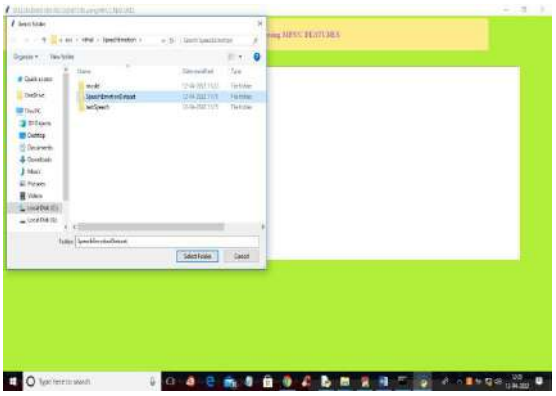

Fig.4.2 selecting and uploading 'Speech Emotion' folder

On the above screen, select and upload the 'Speech Emotion' folder, and then click on the 'Select Folder' button to load the dataset. To get the below screen,

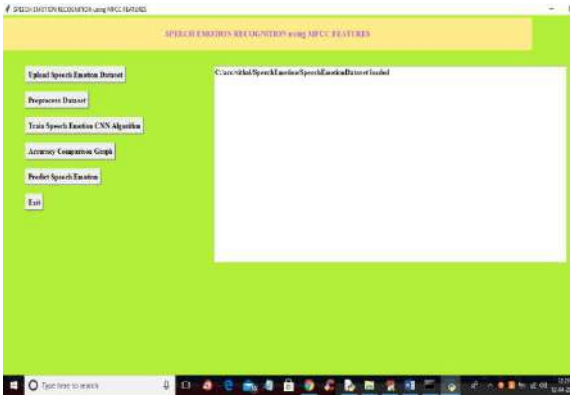

Fig.4.3 Preprocess Dataset

In above screen dataset loaded and now click on 'Preprocess Dataset' to read all audio file and then extract MFCC features and then build X and Y training data and get below output

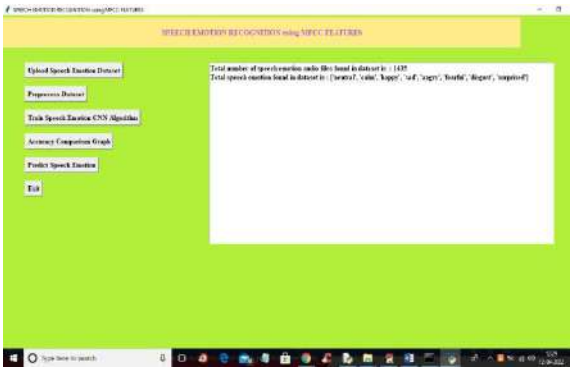

Fig.4.4 dataset processed

In above screen we can see dataset processed and contains 1435 files with 8 different emotions and now click on 'Train Speech Emotion CNN Algorithm' button to train CNN and get below output

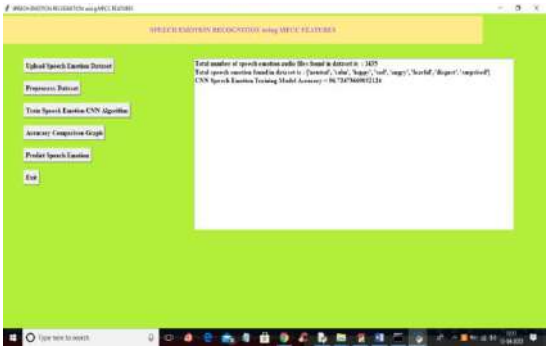

Fig.4.5 Train Speech Emotion CNN Algorithm

In above screen CNN model trained and we got accuracy CNN accuracy as 96% and now click on 'Accuracy Comparison Graph' button to get below graph

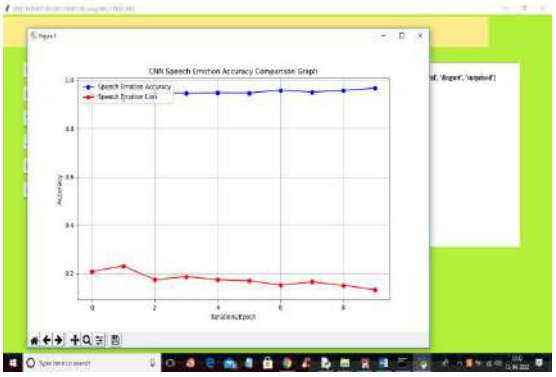

Fig.4.6 Accuracy Comparison Graph

In the above screen, the x-axis represents the training epoch, the y-axis represents accuracy and loss values, and the red line represents loss and the blue line represents the accuracy of CNN. In the above graph, we can see that with each increasing epoch, accuracy increased and loss decreased. Now, close the above graph and then click on the 'Predict Speech Emotion' button to upload an audio file, and then the CNN algorithm will predict the emotion in that file.

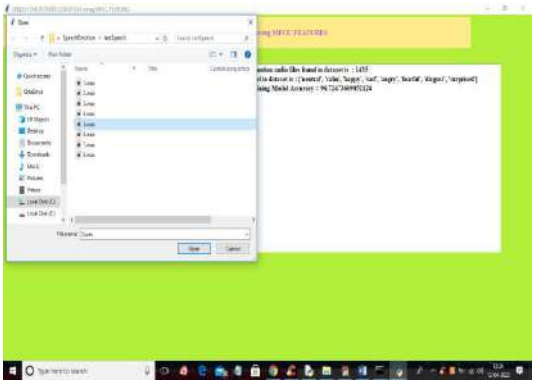

Fig.4.7 selecting and uploading 5.wav file

In above screen selecting and uploading 5.wav file and then click on 'Open' button to load file and get below output

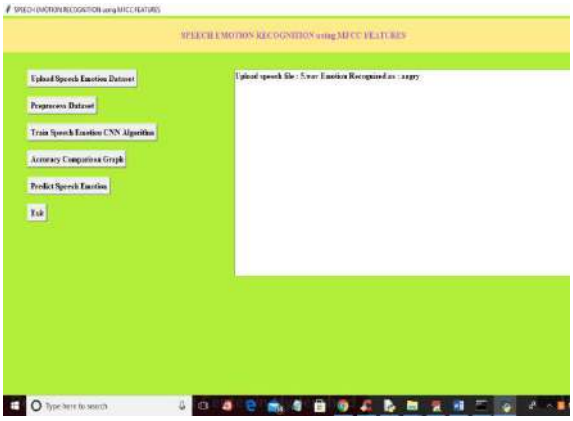

Fig.4.8 Emotion predicted as 'angry'

In above screen in text area we can see emotion predicted as 'angry' and after prediction you can hear audio of that file and now test other files

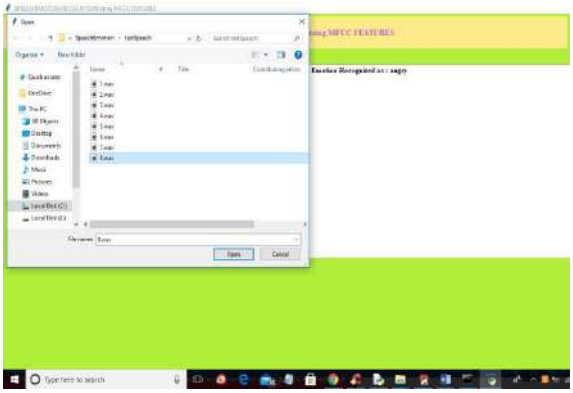

Fig.4.9 uploading 8.wav file

In above screen uploading 8.wav file and get below output

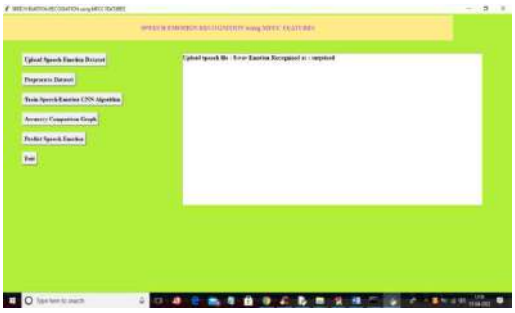

Fig.4.10 audio file emotion detected as surprise

In above screen audio file emotion detected as surprise

V. CONCLUSION

Our work can be expanded to integrate with robots to help them understand the emotional state of their human counterparts, allowing them to have more meaningful conversations with them. It can also be utilised in e-commerce apps like Amazon to make more personalised product recommendations to users based on their browsing history and listening preferences.

REFERENCES

1. Rao, K. Sreenivasa, et al. "Emotion recognition from speech." International Journal of Computer Science and Information Technologies 3.2 (2012): 3603-3607.

2. Yu, Feng, et al. "Emotion detection from speech to enrich multimedia content." Pacific-Rim Conference on Multimedia. Springer, Berlin, Heidelberg, 2001.

3. Utane, Akshay S., and S. L. Nalbalwar. "Emotion recognition through Speech." International Journal of Applied Information Syatems (IJAIS) (2013): 5-8.

4. Manas Jain, Shruthi Narayan, Pratibha Balaji, Bharath K P, Abhijit Bhowmick, Karthik R, Rajesh Kumar Muthu "Speech Emotion Recognition using Support Vector Machine" https://doi.org/10.48550/arXiv.2002.07590

5. Cowie, R., Douglas-Cowie, E., Tsapatsoulis, N., Votsis, G., Kollias, S., Fellenz, W., and Taylor, J. G., Emotion recognition in human-computer interaction, IEEE Signal Processing magazine, Vol. 18, No. 1, 32-80, Jan. 2001.

6. D. Ververidis, and C. Kotropoulos, Automatic speech classification to five emotional states based on gender information, Proceedings of the EUSIPCO2004 Conference, Austria, 341-344, Sept. 2004.

7. Lijiang Chen, Xia Mao, Yuli Xue, Lee Lung Cheng, Speech emotion recognition: Features and classification models, Digital Signal Processing, Volume 22, Issue 6,2012, Pages 1154-1160, ISSN 1051-2004, https://doi.org/10.1016/j.dsp.2012.05.007.