ISSN (Online): 2583-7052

International Journal of Advanced Trends in Engineering and Management (IJATEM)
pp. 195-207

Article Title: Speech Emotion Recognition Using Machine Learning

Speech Emotion Recognition Using Machine Learning

A.S. Sai Puneeth Theja[1], P. Prasanna[2], S. Prathush[3], M. Mounika[4], M. Vinod[5], M. Swetha[6]

[1]Assistant Professor, Department of Electronics and Communication Engineering, Sri Venkatesa Perumal College of Engineering and Technology, Puttur, AP, India.

[2,3,4,5,6] UG Students, Department of Electronics and Communication Engineering, Sri Venkatesa Perumal College of Engineering and Technology, Puttur, AP, India.

ABSTRACT

The task of speech emotion recognition in human-computer interaction is both fascinating and difficult. The practice of attempting to discern affective and emotional states in speech is known as speech emotion recognition. Speech emotion recognition is the process of accurately anticipating a person's emotion from their speech. Many states, such as tone, pitch, expression, behaviour, etc., can be used to forecast an individual's emotion. A select few of them are thought to be able to convey emotion through speaking. With the aid of feature extraction approaches that extract characteristics like MFCC (Mel frequency Cepstral Coefficients), chroma, and Mel spectrogram, the proposed system in the project is able to recognize emotions on dataset. Additionally, the suggested system has real-time emotion recognition capabilities that enables possibility to identify emotions.

Keywords: Speech emotion recognition; Machine learning; Speaker-independent experiment; Classification; Audio emotion recognition

1. Introduction

Emotions are crucial for human communication, expressed through various means like facial expressions and body language. Researchers now focus on understanding emotions through sound signals. This helps in better interaction between humans and computers. Emotions convey feelings effectively, making emotion recognition a popular research area. Now a days, emotion recognition has become a very hot topic for researches. It facilitates smoother communication between people and machines. Emotions can be recognized through facial expressions, body language, and voice. While face-to-face conversations allow easy emotion analysis, it's harder through other channels. Speech emotion recognition (SER) identifies emotions through speech. Humans uniquely express emotions through vocal sounds, like loudness, pitch, and tone. Some universal emotions, such as anger, sadness, and happiness, are easily identifiable. Extracting features from audio signals helps recognize emotions. This is particularly beneficial for physically disabled individuals who struggle to express emotions. Speech emotion recognition is a technology that extracts emotion features from computer speech signals. However , speech is the most effective way to understand what a person wants to convey and humans are interact with each other in various ways such as verbal and non -

ISSN (Online): 2583-7052

International Journal of Advanced Trends in Engineering and Management (IJATEM)
pp. 195-207

Article Title: Speech Emotion Recognition Using Machine Learning

verbal communication .emotions can be of various types like happy, sad, angry, surprise , fear and neutrality that any system can be trained to identify easily. For interaction between human and machine use of speech signal is the fastest and most efficient method. Speech is series sequences of a words of pre -established language and it is an essential medium for communication. Speech technology is a computing technology that empowers an electronic devices to recognizes, to understand spoken words or audio. a lot of machine learning models have been developed and tested in order to classify these emotions carried by speech. And ML algorithms are used to predict the emotions in speech sample, to train the RAVDESS Dataset are used which contains speech samples that were generated by the 24 professional actors. Speech emotion recognition was first developed, its main objective was for the needs of human psychology research. Emotions are not only understood by facial expressions but also with speech. Every speech of human being is associated with an emotion. SER has become an important building block for many smart services systems in areas such as healthcare, smart homes and emergency call centres can use speech emotion analysis to predict to identify the emotions.

1.1 Importance

Speaking up is essential to expressing oneself. The majority of the human body and voice are used in effective communication. One can convey their feelings through hand gestures, body language, tone, and temperament together. Although the verbal portion of communication differs depending on the languages spoken around the world, the non-verbal portion expresses emotions and is probably shared by all. Thus, any cutting-edge technology created to create the impression of a social setting also includes the ability to interpret the emotional context of speech. Advances in emotion detection have a positive effect on many different applications. Among the scientific fields that stand to gain by automating the emotion detection method are neurology, psychiatry, and psychology.

2. Literature Review

In this paper, the extraction author, Girija Deshmukh, Apurva Gaonkar, Gauri Golwalkar, Sukanya Kulkarni, et al. [1], introduced a number of methods; however, the most important and effective algorithm is the one that extracts and improves the GMM, which is best for features utilizing the acoustic speech emotion characteristics of the characteristics like recognition speech signal features. a system that allowed them to hear audio samples of the emotions of irritation, happiness, and melancholy as well as the Short-Term Energy (STE), pitch, energy, and MFCC coefficients. Conclusions and Findings: Train and test sets are carefully separated from the entire Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) dataset. Feature vectors are fed into the multi-class Support Vector Machine (SVM), which outputs a model for each emotion.

ISSN (Online): 2583-7052

International Journal of Advanced Trends in Engineering and Management (IJATEM)
pp. 195-207

Article Title: Speech Emotion Recognition Using Machine Learning

Dave Parry, Christian Poellabauer , John Michael Templeton, Talen Chen, Olayinka Adeleye, Samaneh Madanian, et al.[2]The authors of this research present a variety of extraction techniques, but the most important algorithm is employed in the RNN (Recurrent Neural Network) classifier, as their results are compared to those of CNN (Convolutional Neural Network), SVM (Support Vector Machine), and LR (Logistic Regression).which works well for machine learning-based speech emotion identification feature selection and acaustic feature extraction. A system that allowed them to gather audio samples for the Mel Spectrogram, CHROMA, and MFCC (Mel Frequency Cepstral Coefficient).In happiness, rage, despair, shock, terror, and contempt. Results and Discussion: Train and test sets are carefully separated from the entire Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) dataset. To categorize, a Recurrent Neural Network (RNN) is utilized.

Neethu Sundar Prasad et.al [3]. In this paper, speech emotion detection using machine learning technique is like training a computer to recognize feelings in spoken words by learning from examples, enabling it to classify emotions such as joy, sadness, or anger. A system in which they obtained audio samples extraction authors showed various introduces the to methods but the most significance algorithm is used in in the DCNN, which is less accuracy and low efficiency so move different algorithms in the speech emotion detection using machine learning techniques.

MISSING Likitha and others [4]. In this paper, speech-based human emotion recognition using MFCC is a waveform evaluation of spoken communication to classify the required emotion based on its characteristics such as sound, shape, phoneme, and training. A good number of algorithms have been made from the speech signal in terms of feature extraction and exploration. Acoustic accuracy in communicative kinesics is a property. Feature deletion is the process of deleting a small amount of data. Audio signal used for rear mirroring from each speaker. Most extraction methods are readily available, but a widely used method is the multiplier (MFCC). A feature is a sound representing a speech signal. The process of extracting a small amount of information from a verbal expression that later can be used to operate on each speaker is called feature extraction.

Abhijit Mohanta et al **[5].** In this article, speech emotion detection of speech signal analysis emotions such as anger, fear, happy, neutral of emotional speech signal, where features such as loudness, chord region detection and excitation energy were used in the analysis. They call these functions subsegment functions. The analysis of these emotions uses features such as instantaneous fundamental frequency (FO) using a zero-pass filter (ZFF), signal energy, formant frequencies and dominant frequencies. The study analysis a generation. Characteristics of four different emotional states, rather than **a** classification of emotional states described by the actor. Some signal processing methods, ZFF and STE, were used to locate the current FO and ZCR.

ISSN (Online): 2583-7052

International Journal of Advanced Trends in Engineering and Management (IJATEM)
pp. 195-207

Article Title: Speech Emotion Recognition Using Machine Learning

3. Existing System

The speech emotion recognition system is implemented with a machine learning (ML) model. The implementation steps are comparable to other ML algorithms with additional fine-tuning to improve model performance. The developed model learns from the data provided to it, and all decisions and results that the developed model produces are driven by the data. In the existing system, DCNN is used to predict the accuracy of the model. The flowchart gives a pictorial over view of the process (see figure 1).

3.1 Purpose

The Purpose of this document is speech emotion detection using machine learning algorithms. In detail, this document outlines our project's aim of using machine learning to detect emotions in speech. It will detail the project's overall description, covering user needs, how the product fits into the broader context, and the basic requirements and limitations. Additionally, it will specify the exact requirements and functions necessary for the project, including how the interface should look, what features it must have, and how well it needs to perform.

3.2 Classifiers

Classification is the process of organizing things into different groups based on their characteristics. In machine learning, we use data to train algorithms to recognize patterns and sort new data into categories. For example, email providers use classification to decide if an email is spam or not. It's like finding similarities in data to make predictions about new data. We'll delve into different classification methods and see how text analysis software can understand sentiments in text, like determining if it's positive, negative, or neutral.

3.3 Existing Block Diagram

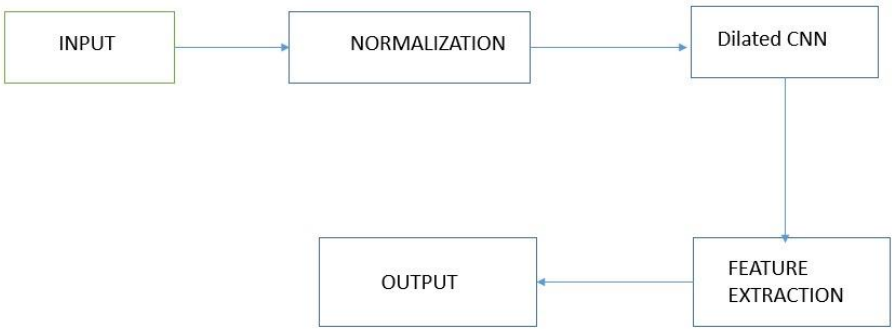

Figure 1: *Existing block diagram*

The first step is data collection, which is of utmost importance. The model developed by RAVDESS Dataset that is used learns from the data provided to it and all the decisions and

ISSN (Online): 2583-7052

International Journal of Advanced Trends in Engineering and Management (IJATEM)
pp. 195-207

Article Title: Speech Emotion Recognition Using Machine Learning

all the decisions and results that the developed model produces are driven by the data. Here RAVDESS (Ryerson Audio Visual Database of Emotional Speech and Song Dataset) means a dynamic, multi-model set of facial expressions and songs, and is a system for classifying different audio speech files into different emotions, and it contains 7356 files and the total size is 24.8 GB and it has 24 professional actors and 12 female actors and 12 male actors. The RAVDESS dataset is a combination of song and speech files, we have a total of 24 actors in both speech and song [13]. 40 in audio files in song files.

The second step is called normalization. It is the process of organizing data into a database. Due to the normalization properties, choosing a suitable normalization algorithm is a difficult task, because the algorithm can affect the classification accuracy. The third phase is often considered the core of an ML project, where the algorithm-based model is developed. Here we think that DILATED CNN (Convolutional Network) is a type of neural network architecture used to process sequential data such as speech signals. Its accuracy is 65.32%. It is called "extended" because it allows the network to gather information from a wider range of input data, skipping some input values. This helps handle input cycles efficiently while maintaining essential functionality. Speech emotion recognition uses advanced CNN to automatically learn patterns and features of speech signals that show different emotions such as happiness, sadness, anger, etc. they are effective in capturing long-range dependencies in the input data and thus can improve the accuracy of emotion recognition systems.

The next step is feature extraction, which is the process of identifying and selecting the most important information or features from a dataset. Current acoustic features commonly used in speech emotion detection include time domain features. Here, an acoustic property means that it is a necessary step in audio signal processing, which is a subfield of signal processing. From this audio data we extracted three main features like MFCC, MEL SPECTROGRAM, and CHROMA.

3.4 MFCC (Mel-Frequency Cepstral Coefficients)

MFCC is one of the popular features used for recognizing the vocal tract mainly used to characterize speakers, from the audio of the speakers, MFCC is a method oof feature extraction of speech and audio analysis. When humans make sounds, it's because of how their vocal cords are shaped. If we know the shape well, we can accurately represent any sound produced. The envelope of the time power spectrum of speech shows the shape of the vocal tract. MFCC is a way to capture this envelope accurately.

In MFCC, we focus on the first 13 coefficients because they capture the main characteristics of the vocal tract's shape, which we call the envelope. We don't need the higher coefficients because they represent smaller details in the sound. Since the envelope is enough to tell the difference between different sounds, we can use MFCC to recognize different sounds, like

ISSN (Online): 2583-7052

International Journal of Advanced Trends in Engineering and Management (IJATEM)
pp. 195-207

different phenomes in speech. MFCC are commonly used in speech and audio processing tasks, and MFCC has been trained to classify eight different emotions.

3.5 MEL Spectrogram

A Mel spectrogram, in the context of speech emotion recognition, is a representation of speech signals that emphasizes frequencies relevant to human hearing. It's derived from the traditional spectrogram, which shows how the frequency content of a signal changes over time. The Mel scale is a perceptual scale of pitches that's based on how humans hear sound.

A Mel spectrogram uses this scale to group frequencies in a way that matches human perception. This means it gives more importance to frequencies that are more distinguishable to the human ear. In speech emotion recognition, a Mel spectrogram is useful because it captures the relevant acoustic features of speech that convey emotions. By analyzing the Mel spectrogram of speech signals, machine learning algorithms can extract features that help classify different emotional states, such as happiness, sadness, anger and so on.

3.6 Chroma

In speech emotion recognition using machine learning, "chroma" refers to a feature extraction technique that focuses on the pitch content of the speech signal. Specifically, chroma features represent the distribution of energy in different pitch classes, and it represents total content of a musical audio signal in a condensed form. Chroma features are useful in capturing tonal aspects of speech, such as intonation and melody, which are important cues for conveying emotions.

By the chroma features extracted from speech signals, machine learning algorithms can identify patterns related to different emotional states, such as happiness, sadness, or excitement. Overall, chroma features provide valuable information for understanding the emotional content of speech and can be used as input features for machine learning models in speech emotion recognition systems.

The final step is to evaluate the output is represented in a numerical value each corresponds to either of expressions like angry, surprise, joy, sad, fear, Neutral, Disgust etc.

4. Proposed Work

The proposed system can focus on important parts of the face and achieve improvements over previous models. From the experiments, it was proved that different emotions seem to be sensitive to different parts of the face. A visualization method to highlight the salient regions of the face images were deployed which highlights the salient regions of face images which are considered the crucial parts in detecting different facial expressions.

ISSN (Online): 2583-7052

International Journal of Advanced Trends in Engineering and Management (IJATEM)
pp. 195-207

Article Title: Speech Emotion Recognition Using Machine Learning

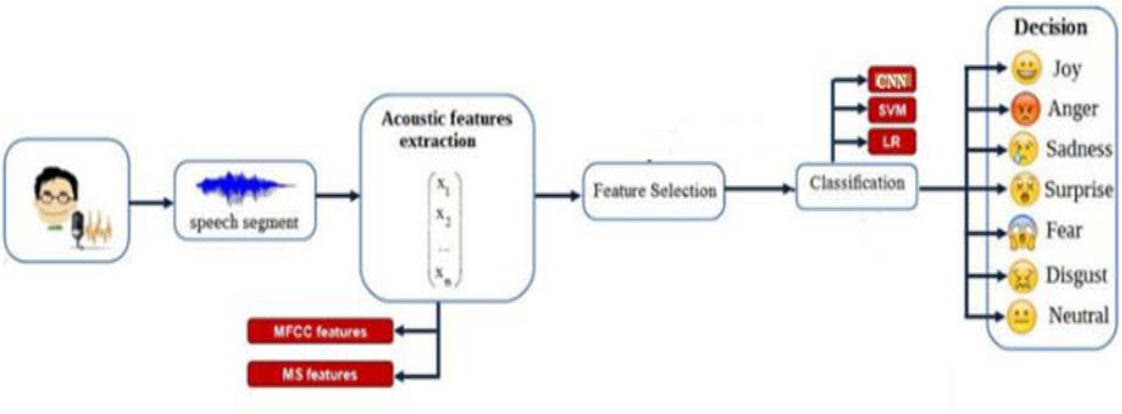

Figure 2: *Proposed block diagram*

Speech Emotion Recognition 'SER' involves understanding emotions in spoken words by using different classifiers and different methods feature is developed. The project utilizes Mel-frequency cepstral coefficients 'MFCC' and modulation spectral (MS) features extracted from speech signals and used to train different classifiers. Feature selection 'FS' is applied to identify the most relevant feature subset for training classifiers. We tried to determine which feature are most relevant to differentiate each emotion A recurrent neural network 'RNN' classifier is used to classify seven emotions initially, performance are compared to SV (Support vector machine) ,CNN (Convolutional Neural Network) ,LR(Logistic Regression). These machine learning techniques main aim to enhance the accuracy of emotion classification in spoken audio signals. CNN (Convolutional Neural Network) is a type of Artificial Neural Network.

In SER CNNS are used to automatically learn features from speech signals that corresponds to different emotions. The goal is to train a CNN model to identify and distinguish different voices, enabling the system to recognize and understand human speech accurately.) SVM (Support Vector Machine) is a Supervised Machine Learning Algorithm used for classification tasks. SVM can be trained on features extracted from speech signals such as (MFCC, MEL SPECTROGRAM) to classify the emotions expressed in the speech.3) LR (Logistic Regression) can be used to classify whether a given speech signals correspond a specific emotion or not While LR is simpler compared to SVM or CNN.

A Recurrent Neural Network is a model that is trained to process and convert a sequential data input into a specific sequential data output. RNN is used to predicts the emotions. RNN is a powerful models for time –series classification Data arranged in sequence. MLP (Multi Layer Perceptron) model is considered to the suitable model for speech emotion recognition. This is because, MLP classifier is best suitable for complex structures datasets when compared to other models like SVM, CNN etc.MLP classifier is a Neural Network to solve classification prediction problems.

ISSN (Online): 2583-7052

International Journal of Advanced Trends in Engineering and Management (IJATEM)
pp. 195-207

Article Title: Speech Emotion Recognition Using Machine Learning

4.1 Machine Learning

Machine learning algorithms are categorized into supervised, unsupervised and semi-supervised learning. Supervised learning is learning from data that provides corrective information to the algorithm. Unsupervised learning is the learning of patterns without training data. Whereas semi-supervised learning is the learning with partially label data or by receiving a reward from the environment. Fluid mechanisms were traditionally concerned with big data and currently, fluid mechanisms are beginning to tap into the full potential of the powerful methods [3].

Supervised learning is a widely used machine learning method. It includes spam classifiers of email, face-recognizers over images and medical diagnosis systems for patients. Gradient-based optimization algorithms are being used by the deep learning systems. Major trend is the growing concern with the environment in which a machine-learning algorithm operates. New machine-learning methods that are capable of working collaboratively with humans to jointly analyzes complex data sets. Recent progress in machine learning had driven the development of new algorithms. Adoption of data-intensive methods can be found throughout science, technology, and commerce which leads to more evidence-based decision making [2].

4.2 Feature Extraction

The next step involves extracting the features from the audio files which will help our model learn between these audio files. For feature extraction we make use of the LIBROSA library in python which is one of the libraries used for audio analysis, when we do Speech Recognition tasks, MFCCs are the state-of-the-art feature since it was invented in the 1980sThis shape determines what sound comes out. If we can determine the shape accurately, this should give us an accurate representation of the phoneme being produced. The shape of the vocal tract manifests itself in the envelope of the short time power spectrum, and the job of MFCCs is to accurately represent this envelope.

5. Methodology

There are four primary processes in our SER system. The collection of voice samples comes first. After the features are extracted, a second features vector is created. The next thing we did was try to figure out which features are most important for differentiating between each emotion. These features are added to the recognition machine learning classifier.

The audio that was utilized as the input for the MLP classifier model is processed by the suggested method to extract characteristics such as chroma, Mel spectrogram, and Mel-Frequency Cepstral Coefficients (MFCC). This suggested system trains and predicts emotions like calm, happiness, fearfulness, and disgust using the MLP classifier. The suggested system's prediction accuracy is maximized by the use of hyperparameter adjustment. A neural network called the MLP classifier is employed to solve classification issues. Since it is a supervised

ISSN (Online): 2583-7052

International Journal of Advanced Trends in Engineering and Management (IJATEM)
pp. 195-207

Article Title: Speech Emotion Recognition Using Machine Learning

learning algorithm, it requires a dependent variable (the target variable) and one or more independent variables.

Mel-scaled spectrograms and MFCCs are commonly used in speech emotion identification and sound classification applications. These characteristics somewhat resemble the way that humans naturally absorb sound frequencies. Mel Scale is created by converting the typical frequency of human speech using MFCC. A descriptor that condenses the tonal content of a musical audio source is called a chroma or chromogram feature. Chroma characteristics are mostly used to combine all of the spectral data associated with a pitch into a single coefficient.

6. Results and Discussion

The simulation has been done on the Anaconda, A Python distribution using Python's Jupyter Notebook tool. The steps to execute the simulation is as mentioned below. Furthermore, the results for various scenarios have also mentioned.

Step1: Provide the Sample Audio as the Input.

Step2: From the audio file, the Spectrogram and wave form are plotted.

Step3: Extract the Mel Frequency Cepstral Coefficients (MFCC) using the LIBROSA ,a python library.

Step 4: Constructing layers to train the dataset, remixing the data, and dividing it into train and test.

Step5: Using the trained data to predict the emotion of a human voice (Sample No -Predicted Value-Actual Value).

SCENARIO 1: In this Initial scenario, taking a Female sample audio with a happy emotion as the input

Input:

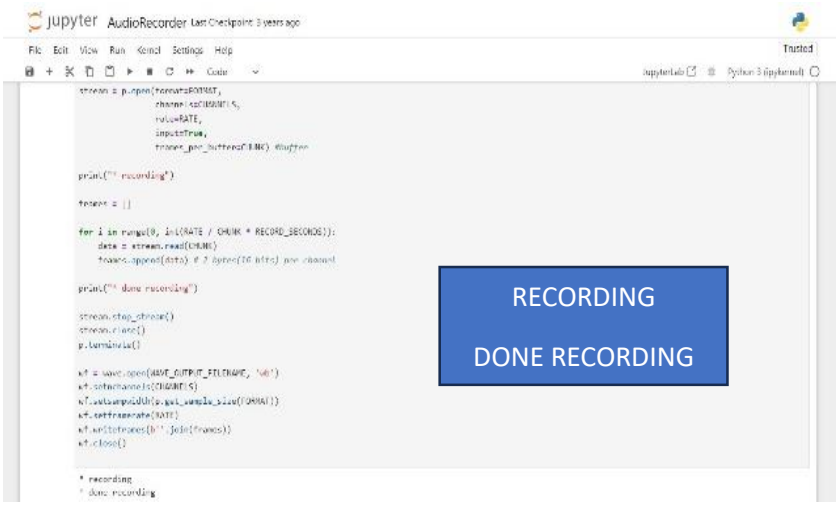

Figure 3: *Recording audio for Scenario-1*

ISSN (Online): 2583-7052

International Journal of Advanced Trends in Engineering and Management (IJATEM)
pp. 195-207

Article Title: Speech Emotion Recognition Using Machine Learning

Output:

Get the output in the sample audio as the above scenario 1. The output is in the form of a text and also in the form of voice

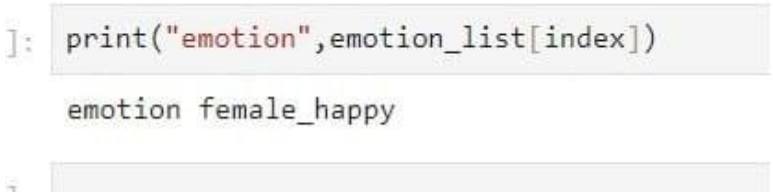

Figure 4: *Output for Scenario-1*

SCENARIO 2: In the next scenario, taking a Female sample audio as the input

Input:

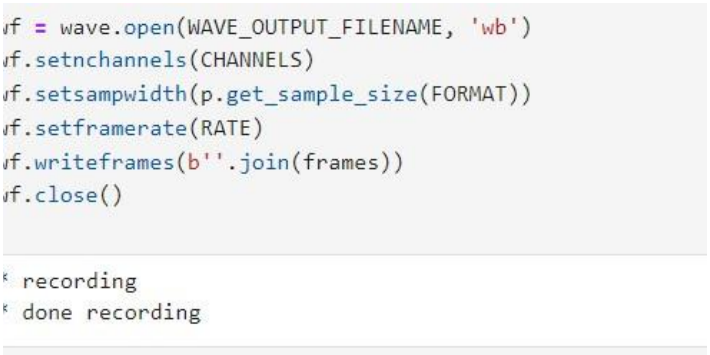

Figure 5: *Recording audio for Scenario-2*

Output: The output is displayed as a female_sad, which indicates that the given sample is a female voice with sad emotion. The output is as shows in the figure.6

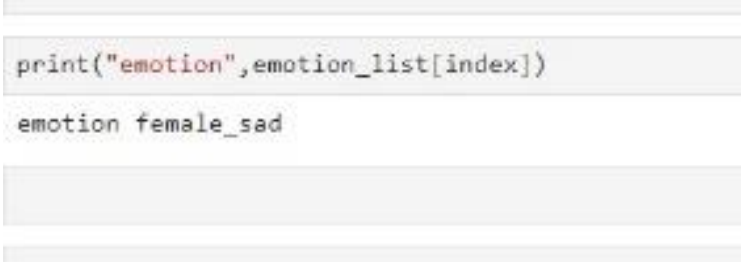

Figure 6: *Output for Scenario-2*

ISSN (Online): 2583-7052

International Journal of Advanced Trends in Engineering and Management (IJATEM)
pp. 195-207

Article Title: Speech Emotion Recognition Using Machine Learning

SCENARIO 3: In the third scenario, taking a Female sample audio as the input

Input:

Figure 7: *Recording audio for Scenario-3*

Output:

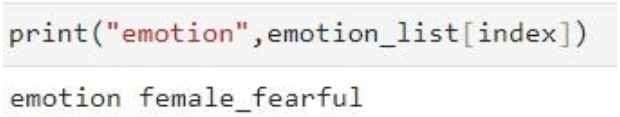

Figure 8: *Output for Scenario-3*

SCENARIO 4: In the fourth scenario, taking a sample audio as the input

Input:

Figure 9: *Recording audio for Scenario-4*

ISSN (Online): 2583-7052

International Journal of Advanced Trends in Engineering and Management (IJATEM)
pp. 195-207

Article Title: Speech Emotion Recognition Using Machine Learning

Output:

```
print("emotion",emotion_list[index])

emotion female_angry
```

Figure 10: *Output for Scenario-4*

SCENARIO 5: In the fifth scenario, taking a male sample audio as the input
Input: The recording for the male voice is taken from the souce by running the python code as shown in the figure 11

```
wf = wave.open(WAVE_OUTPUT_FILENAME, 'wb')
wf.setnchannels(CHANNELS)
wf.setsampwidth(p.get_sample_size(FORMAT))
wf.setframerate(RATE)
wf.writeframes(b''.join(frames))
wf.close()

* recording
* done recording
```

Figure 11: *Recording audio for Scenario-5*

Output:

Get the output in the sample audio as the above scenario 5.

```
print("emotion",emotion_list[index])

emotion male_sad
```

Figure 12: *Output for Scenario-5*

From all the five scenarios it is evident that the given audio sample to the code generate and detect the emotion of a person with in its scope.

7. Conclusion

In conclusion, speech emotion recognition using machine learning is a promising technology that can understand and classify emotions from spoken words. By analyzing features like tone,

pitch, and expression, machine learning algorithms can accurately identify emotions like happiness, sadness, anger, and more.

This technology has the potential to enhance various applications, from customer service to mental health support, by providing valuable insights into people's emotions through their speech patterns. The Speech emotion recognition (SER) is a field that recognizes human emotions through the speech. A correct and precise database where the actors' voice is clear and noise free is ideal. An overview of SER methods is discussed for extracting audio features from speech sample, various classifier algorithms are used to recognise the emotion.

Various features are used to recognise emotions but feature extraction using MFCC seemed to have an important role in recognizing emotions through the speech. In this study, we saw various classifiers being used for classification. Whereas choosing the proper and best classifier is very important step in SER. The accuracy of the Speech emotion recognition system is dependent upon the Database, the features extracted from the databases and a classification model (algorithm) used to classify the emotions.

References

1. "Speech based Emotion Recognition using Machine Learning," Institute of Electrical and Electronics Engineers, March 2019, Girija Deshmuck, Apurva Gaonkar, Gauri Golwalkar, and SukanyaKulkarni.
2. Talen Chen, Samanet Madanian (2022). Using machine learning to recognize emotions in speech An Organized Assessment Auckland Department of Software Engineering and Computer Science.
3. Neethu Sundar Prasad, Institute of Electrical and Electronics Engineers, "Speech Emotion detection using machine learning technique," June 2019.
4. "Speech Based Human Emotion Recognition Using MFCC," Institute of Electrical and Electronics Engineers, March 2017, Sri Raksha R. Gupta, M.S. Likitha, A. Upendra Raju, and K. Hasitha.
5. Institute of Electrical and Electronics Engineers, "Speech Emotion Recognition from Speech Signal," Abhijit Mohanta, Vinay Kumar Mittal.November of 2017.
6. Sri Raksha R. Gupta, M.S. Likitha, A. Upendra Raju and K. Hasitha "Speech Based Human Emotion Recognition Using MFCC", Institute Of Electrical And Electronics Engineers, March 2017.
7. Ye Sim Ülgen Sonmez, Asaf Varol, "New Trends in Speech Emotion Recognition", Institute of Electrical and Electronics Engineers, June 2019.
8. Pavol Harár , Radim Burget , Malay Kishore Dutta, "Speech emotion recognition with deep learning", Institute Of Electrical And Electronics Engineers, Feb. 2017.
9. Esther Ramdinmawii, Abhijit Mohanta, Vinay Kumar Mittal, "Emotion recognitionfrom speech signal", Institute Of Electrical And Electronics Engineers, Nov. 2017.
10. Peng Shi, "Speech Emotion Recognition Based on Deep Belief Network", Institute Of Electrical And Electronics Engineers, March 2018.