International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 08 Issue: 10 | Oct 2021 www.irjet.net p-ISSN: 2395-0072

Speech Emotion Recognition using Machine Learning

Sayeed Anwar Sheikh[1], Amaan Syed[2], Osama Ansari[3], Aarish Fakih[4], Anand Bali[5], Ajitkumar Khachane[6]

[1-4]Student, Dept. of Computer Engineering, M.H Saboo Siddik College of Engineering, Maharashtra, India.
[5]Professor, Dept. of Computer Engineering, M.H Saboo Siddik College of Engineering, Maharashtra, India.
[6]Professor, Dept. of Information Technology Engineering, Vidyalankar Institute of Technology, Maharashtra, India.

Abstract - *Communication is essential for clearly expressing one's thoughts and ideas. Speech is the most favored and powerful means of communication in human beings. The Internet of Things (IoT) age is quickly progressing, bringing more sophisticated technologies into everyday use. Simple wearables and widgets to complex self-driving vehicles and automated systems are all examples of these applications. Intelligent apps are interactive and require minimal human effort to use, and rely on voice-based input for the most part. As a result, these computer programs must be able to fully grasp human speech. Gender, age, language, and mood are all things that a speech percept might convey about the speaker. Several current voice recognition algorithms that are utilized in IoT applications have been combined with an emotion detection system to assess the speaker's emotional state. In many respects, the performance of the emotion detection system may have a significant impact on the entire performance of the IoT application, and it can give numerous benefits over the functions of these apps. This study offers a speech emotion detection system that outperforms previous systems in terms of data, feature selection, and technique, with the goal of more correctly categorizing speech percepts based on emotions.*

Key Words*: *IoT, RAVDESS, SVM, MLP, MFCC, Speech, voice-based, Emotion, SER, Machine Learning.

1.INTRODUCTION

Speech, tone, pitch, and other features of the human vocal system are used to communicate information and context. As human-machine interactions progress, it will be necessary to strengthen the results of such encounters by giving computer and machine interfaces the capacity to detect the speaker's emotion. Today, a significant amount of money and effort is being invested in the creation of artificial intelligence and smart robots, all with the goal of making human life easier. The advancements of AI have engendered to several technologies involving Human-Computer Interaction (HCI) [1]. Aiming to develop and improve HCI methods is of paramount importance because HCI is the front-end of AI which millions of users experience. Some of the existing HCI methods involve communication through touch, movement, hand gestures, voice and facial gestures [1]. Speech Emotion Recognition (SER) attempts to deduce a speaker's underlying emotional state from his or her voice. Throughout the last several years, there has been a surge in research interest in this field. People's emotions can be detected in a variety of settings, including robot interfaces, audio surveillance, web-based E-learning, commercial applications, entertainment, banking, contact centres, and computer games. Information regarding students' emotional states can help concentrate classroom orchestration or E-learning on improving teaching quality. A teacher, for example, might utilize SER to select which subjects to teach and must be able to create techniques for regulating emotions in the classroom. As a result, the emotional condition of the students should be taken into account in the classroom. The study in a recent article foresees that by 2022, about 12% of all IoT applications would fully function based on voice commands only [2].

Due of its complexity, SER is difficult to integrate with the other components. Furthermore, to be considered intelligent, a computer system must be able to replicate human behaviour. The capacity to change discussions based on the emotional state of the speaker and the listener is a noteworthy feature of human nature. Speech emotion detection is a classification issue that may be handled with a variety of machine learning methods. The goal of this study is to create a system that can automatically recognize eight different emotions from a person with a neurological disorder's speech. This goal may be achieved by using two databases, the RAVDESS dataset and a custom local dataset, to train a neural network. This project goes into the many approaches and tests used to create a Speech Emotion Detection system in great depth.

This paper aims at designing a system which detects the emotion by recognising the tone of the speech. The system detects the speech and compares it with the given dataset and gives the matching emotion as output. This paper is divided into various sections. Page 1 contains Introduction and Literature Survey, page 2 contains Proposed System, Page 3 contains Methodology and Conclusion.

2. LITERATURE SURVEY

Cao et al. [3] proposed a system that considered that the emotion expressed by humans are mostly a result of mixed feeling. As a result, they proposed an enhancement to the SVM algorithm that takes into account mixed signals and selects the most dominant one. A ranking SVM algorithm was selected for this objective. The ranking SVM applies all of the

predictions from the individual binary classification SVM classifiers, also known as rankers, to the final multi-class issue. Their system obtained a 44.40 percent accuracy using the ranking SVM algorithm.

Chen et al. [4] developed a system that had improvements in the pre-processing stage. Fisher and Principle Component Analysis (PCA) were utilized as preprocessing methods in conjunction with SVM and ANN as classification algorithms. They ran four trials, each with a distinct pre-processing and classifier algorithm combination. The Fisher technique was employed in the first experiment to choose features for a multi-level SVM classifier (Fisher + SVM). The second experiment included utilizing Principle Component Analysis (PCA) for the SVM classifier (PCA + SVM) to decrease feature dimensionality. The Fisher method was applied over the ANN model in the third experiment (Fisher Plus ANN). Finally, PCA was used before employing ANN to classify the data (PCA + ANN). Two significant findings were drawn from these studies. For starters, reducing dimensionality increases the system's performance.

Second, in the case of emotion identification, the SVM classifier method performs better than the ANN algorithm. Using Fisher for dimensionality reduction and SVM for classification, the winning experiment achieved an accuracy of 86.50 percent.

In the Nwe et al. [5] system, a subset of features, similar to the Mel Frequency Cepstral Coefficients (MFCC), was used. To identify emotions in speech, they utilized the Log Frequency Power Coefficients (LFPC) over a Hidden Markov Model (HMM). Because they utilized a dataset that was only accessible to them, their work isn't publicly available. They argue, however, that utilizing the LFPC coefficients rather of the MFFCC coefficients improves the model's accuracy significantly. Their model's average categorization accuracy is 78 percent, with the highest accuracy being 96 percent.

Rong et al. [6] proposed an innovative way to improve the accuracy of existing models. To decrease the amount of characteristics, computer scientists have traditionally used different pre-processing methods. This new method, on the other hand, expanded the amount of characteristics utilized for categorization. They claimed to have classified auditory percepts in the Chinese language using a tiny dataset, however they don't specify which characteristics they utilized. They did say, however, that none of their features are language dependent. They obtained an accuracy of 82.54 percent using a large number of features and an ensemble random forest method (ERFTrees).

Narayanan [7], in his work, proposes a system that uses a more real-world dataset. Data was gathered from a contact center for his project, and he used a binary classification system with just two emotions: joyful and furious. Over the KNN algorithm, the researchers utilized a variety of characteristics, including auditory, lexical, and other language-based features. Furthermore, this study was performed especially for the contact center industry and compared male and female consumers. In male and female consumers, accuracy values improved by 40.70 percent and 36.40 percent, respectively.

3. PROPOSED SYSTEM

A Machine Learning (ML) model is used to build the voice emotion detection system. The implementation stages are similar to those for any other machine learning project, with extra fine-tuning methods to improve the model's performance. The flowchart depicts a visual representation of the procedure. The gathering of data is the first and most important stage. The first step in implementing the Speech Emotion Recognition system is to collect audio samples under different emotional categories which can be used to train the model. The audio samples are usually wav or mp3 files. The model being built will learn from the data given to it, and the data will drive all of the choices and outcomes that a completed model will generate. Feature engineering is the second phase, which consists of a set of machine learning tasks that are run over the gathered data. Several data representation and data quality problems are addressed by these methods. The third stage, when an algorithmic-based model is created, is generally regarded the heart of an ML project. This model learns about the data and trains itself to react to any new data it encounters using a machine learning technique. The last stage is to assess how well the constructed model works. To evaluate the performance of various algorithms, developers often repeat the processes of creating a model and assessing it. The results of the comparisons aid in selecting the best suitable machine learning method for the issue.

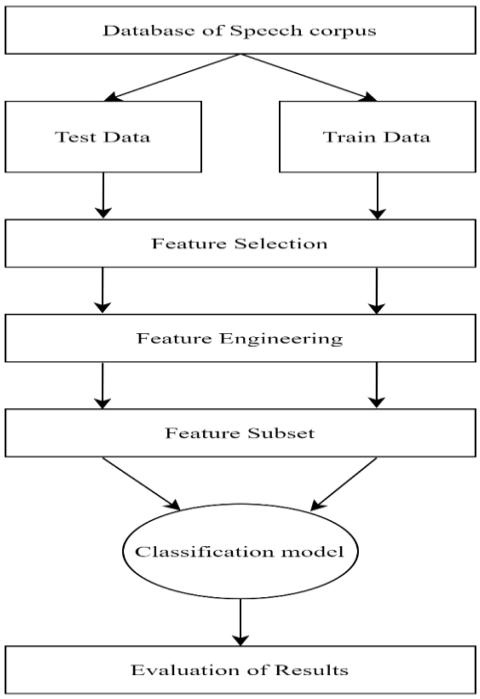

Chart -1: Flow of Implementation

4. METHODOLOGY

The initial stage in this suggested system is to train the model using a RAVDESS dataset and the MLP (Multi-Layer Perceptron) classifier as a machine learning method [9]. The system has two modes of operation: real-time input and input from a file that is already there. The suggested method collects characteristics from audio such as Mel-Frequency Cepstral Coefficients (MFCC), chroma, and mel spectrogram, which are then fed into the MLP classification model.

This suggested system use the MLP classifier to train and predict emotions such as calm, pleasure, fear, and disgust. The suggested system's prediction accuracy is raised to its maximum limit by employing hyperparameter tweaking.

A Neural Network called an MLP classifier is used to solve classification issues. It's a supervised learning method, which means it requires independent and dependent variables (target variable). In the domains of sound categorization and speech emotion identification, mel-scaled spectrograms and MFCCs are frequently used. These features mimic to a certain extent the reception pattern of sound frequency intrinsic to a human [8]. The MFCC method is used to convert the frequency of human speech into Mel Scale. The chroma, also known as a chromogram, is a description that describes the tonal content of a musical audio stream in a reduced form. The primary goal of chroma features is to combine all spectral information for a particular pitch into a single coefficient.

Furthermore, the dataset used to train the model is only a small portion of the RAVDESS dataset, which contains 24 actors and only a handful of the emotions. However, if we utilize the full dataset and the entire spectrum of emotions from the RAVDESS dataset, the model's accuracy should potentially improve.

The suggested method was shown to be 78.65 percent accurate in predicting the aforementioned emotions. The suggested system's real-time functionality was achieved by using the pyaudio package, which captures audio and extracts and predicts characteristics from it.

The suggested system can also predict emotions from a recorded audio file after performing feature extraction and prediction.

The user is given three alternatives to choose from in the proposed system (figure-2):

1. Create and Train Model
2. Record and predict emotion
3. Predict emotion on a specified audio file

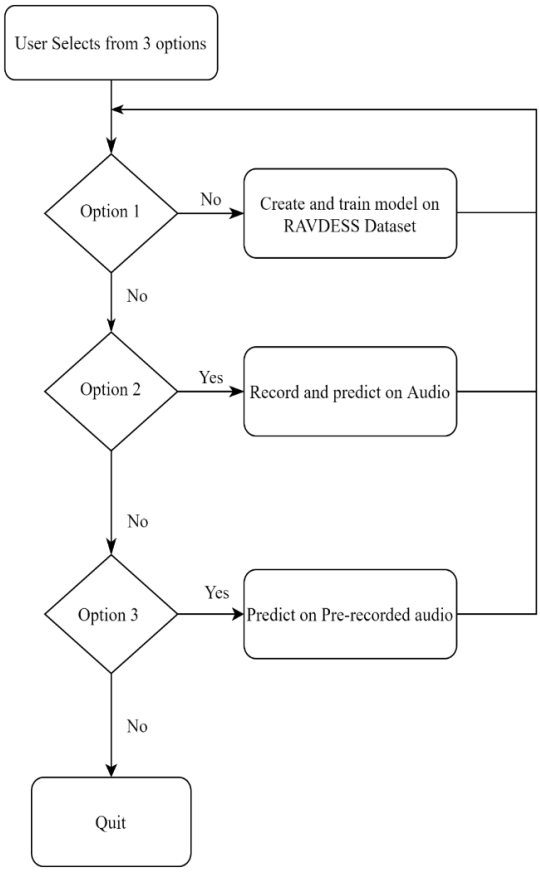

Fig -1: Concept Diagram Speech Emotion Recognition [10]

Sample paragraph Define abbreviations and acronyms the first time they are used in the text, even after they have been defined in the abstract. Abbreviations such as IEEE, SI, MKS, CGS, sc, dc, and rms do not have to be defined. Do not use abbreviations in the title or heads unless they are unavoidable.

After the text edit has been completed, the paper is ready for the template. Duplicate the template file by using the Save As command, and use the naming convention prescribed by your conference for the name of your paper. In this newly created file, highlight all of the contents and import your prepared text file. You are now ready to style your paper.

5. CONCLUSIONS

The new age of automation has begun as a result of the growing growth and development in the fields of AI and machine learning. The majority of these automated gadgets are controlled by the user's vocal commands. Many benefits may be developed over current systems if, in addition to identifying words, the computers also understand the speaker's mood (user). Computer-based instructional applications, automated contact center dialogues, a diagnostic tool for treatment, and an automatic translation system are some of the uses of a voice emotion detection system.

The stages for developing a speech emotion recognition system were described in depth in this thesis, and several tests were conducted to understand the effect of each step. It was initially difficult to build a well-trained model due to the low amount of publicly accessible speech databases. Then, in previous studies, many new methods to feature extraction had been suggested, and choosing the optimal approach required doing several trials. Finally, learning about the strengths and weaknesses of each classifying algorithm in terms of emotion detection was part of the classifier selection process. When comparing a single feature to an integrated feature space, the results show that an integrated feature space produces a higher recognition rate.

The suggested project may be further modeled in terms of efficiency, accuracy, and usefulness for future improvements. The model may be expanded to detect sensations such as sadness and mood swings in addition to emotions. Therapists may utilize such systems to keep track of their patients' mood fluctuations. Incorporating a sarcasm detection system is a difficult result of developing robots with emotion. Sarcasm identification is a more difficult issue than emotion detection since sarcasm cannot be detected just by listening to the speaker's words or tone. To detect sarcasm, a sentiment detection utilizing language may be combined with speech emotion recognition. As a result, numerous applications of a speech-based emotion detection system will develop in the future.

REFERENCES

[1] Soegaard, M. and Friis Dam, R. (2013). The Encyclopedia of Human-Computer Interaction. 2nd ed.

[2] Gartner.com. (2018). Gartner Says 8.4 Billion Connected. https://www.gartner.com/newsroom/id/3598917.

[3] H. Cao, R. Verma, and A. Nenkova, “Speaker-sensitive emotion recognition via ranking: Studies on acted and spontaneous speech,” Comput. Speech Lang., vol. 28, no. 1, pp. 186–202, Jan. 2015.

[4] L. Chen, X. Mao, Y. Xue, and L. L. Cheng, “Speech emotion recognition: Features and classification models,” Digit. Signal Process., vol. 22, no. 6, pp. 1154–1160, Dec. 2012.

[5] T. L. Nwe, S. W. Foo, and L. C. De Silva, “Speech emotion recognition using hidden Markov models,” Speech Commun., vol. 41, no. 4, pp. 603–623, Nov. 2003.

[6] J. Rong, G. Li, and Y.-P. P. Chen, “Acoustic feature selection for automatic emotion recognition from speech,” Inf. Process. Manag., vol. 45, no. 3, pp. 315–328, May 2009.

[7] S. S. Narayanan, “Toward detecting emotions in spoken dialogs,” IEEE Trans. Speech Audio Process., vol. 13, no. 2, pp. 293–303, Mar. 2005.

[8] D. Issa, M. F. Demirci and A. Yazici “Speech emotion recognition with deep convolutional neural networks” 2020, Biomedical Signal Processing and Control.

[9] A. A. A. Zamil, S. Hasan, S. M. Jannatul Baki, J. M. Adam and I. Zaman, "Emotion Detection from Speech Signals using Voting Mechanism on Classified Frames," 2019 International Conference on Robotics,Electrical and Signal Processing Techniques (ICREST), 2019, pp. 281-285.

[10] https://www.irjet.net/archives/V8/i6/IRJET-V8I6496.pdf

[11] https://www.frontiersin.org/articles/10.3389/fcomp.2020.00014/full