iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 11 **Issue:** V **Month of publication:** May 2023

DOI: https://doi.org/10.22214/ijraset.2023.51575

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 11 Issue V May 2023- Available at www.ijraset.com

Speech Emotion Detection Through Live Calls

S. Shreya[1], P. Likitha[2], G. Saicharan[3], Dr. Shruti Bhargava Choubey[4]
[1, 2, 3]*Student,* [4]*Dean-Innovation, Associate Professor, Sreenidhi Institute of Science and Technology (SNIST), Affiliated to JNTUH, ECEDepartment, Ghatkesar, Telangana, India*

Abstract: ***Speech emotion recognition is a popular study area right now, with the goal of enhancing human-machine connection. Most of the research being done in this field now classifies emotions into different groups by extracting discriminatory features. Most of the work done nowadays concerns verbal expressions used for lexical analysis and emotion recognition. In our project, emotions are categorized into the following categories: angry, calm, fearful, happy, and sad. Speech Emotion Recognition, often known as, SER, is a technology that takes advantage of the fact that tone and pitch in a speech frequently convey underlying emotions. The approach to assessing or anticipating a speaker's gender and emotions from their speech has been given in the proposed work. By graphing the waveform and spectrogram, convolutional neural networks are used to evaluate or predict gender and emotions. A CNN model is created using the input of 12162 samples to ordering to identify the emotions present in the speech. In our study, the suggested model's overall accuracy is calculated using only one feature, the MFCC from the speech, and the 4 datasets (RAVDESS, SAVEE, CREAMA-D, and TESS). The accuracy is first calculated for each emotion and gender before the overall accuracy is discovered.***

Keywords: Classification, Emotion recognition, convolution neural networks, lexical analysis, Support vector machine, etc.

I. INTRODUCTION

A speech signal is one of the quickest and most natural ways for humans to communicate. The use of speech signals for human-machine interaction is the quickest and most efficient method. All available senses are used by the human natural ability to maximize awareness of a received message. Emotional detection is a difficult task for machines, but it is natural for humans. As a result, knowledge about emotions is used by an emotion recognition system to improve communication between machines and humans. The female or male speaker's emotions are discovered through speech emotion recognition. Some of the investigated speech are linear prediction cepstrum coefficient (LPCC), fundamental frequencies, and Mel+ frequency cepstrum coefficient (MFCC). These characteristics serve as the foundation for speech processing. It is unclear which speech features are more useful in differentiating between various emotions, which makes emotion recognition from speakers' speech very difficult.

The presence of different speaking rates, styles, sentences, and speakers introduces accosting variability, which affects the features of speech. It is challenging to distinguish between different parts of an utterance that express different emotions because they are expressed in different ways in the spoken word for each corresponding emotion. The expression of emotion is influenced by the speaker's culture and environment, which creates another issue because there is variation in speaking style due to environmental and cultural differences. Transient and long-term emotions are two types of emotions, and it is unclear which type is detected by the recognizer. Emotions recognized in speech may be speaker-independent or speaker dependent. For classification, various classifiers such as K-nearest neighbors (KNN), Support vector machine (SVM), CNN, and others are available. In the second section of this paper, a brief introduction to speech emotion recognition is provided, along with a description of the speech emotion recognition system block diagram. Various work has been done on various datasets, so some of the existing datasets, as well as modeling of emotional speech and different types of speech, are covered in the third section. The fourth section provides brief details about various feature extraction mechanisms for speech emotion recognition, followed by a review of the classification section. This section has covered KNN, SVM, CNN, recurrent neural networks, and other topics. The sixth section provides a quick overview of the use of deep learning for speech emotion recognition.

II. METHODOLOGY

The voice emotion recognition system is created with a machine learning (ML) model. Similar to every other ML project, the implementation process includes additional fine-tuning processes to improve the model's performance. A visual overview of the procedure is provided by the flowchart. Data collection is the initial phase, which is crucial. All of the decisions and outcomes that a built model will make are informed by the data, and the model is learning from the data that is being presented to it. The second step, referred to as feature engineering, consists of a number of machine learning operations that are carried out on the gathered data. Many problems with data representation and data quality are addressed by these approaches.

The development of an algorithmic-based model takes place in the third step, which is frequently regarded as the project's heart. This model trains itself to react to any new data it is exposed to using an ML algorithm to learn about the data. Evaluation of the developed model's performance is the last phase. The process of creating a model and analyzing it is frequently repeated by developers in order to assess the effectiveness of various algorithms. Results of comparisons aid in selecting the ML method most appropriate to the issue.

A. Data Collection

The collection of audio samples under various emotional categories that can be used to train the model is the first stage in putting the Speech Emotion Recognition system into practice. The audio samples are typically wav or mp3 files that are freely downloadable by the general audience. The subsequent stages are described in relation to the tests conducted using the TESS dataset.

B. Data Visualization

The problem and the type of solution to be developed are better understood through the data's visual representation. Many techniques for displaying the data include clustering, the distribution of classes, the number of instances within each category, the spread of the data, and the correlation between the attributes. Statistical functions for data visualization are available in Python and R.

C. Data Preparation

It's time to get the data ready for processing once you've done some data analysis utilizing different visualizations. The steps involved in data preparation include correcting difficulties with quality, standardization, and normalization. The data are initially checked for problems including missing values, outliers, erroneous data, and duplicate data. The dataset had no invalid, duplicate, or missing values. With the help of the outlier analysis, it is possible to see how many outliers there are for each feature as well as how the mean value of the feature varies with and without the outliers. This will make it easier to determine whether the outliers are really outliers or whether they influence decision-making. The data was then subjected to normalization because the raw data had been collected on a different scale. After standardization, the range of all feature values is now 0 to 1.

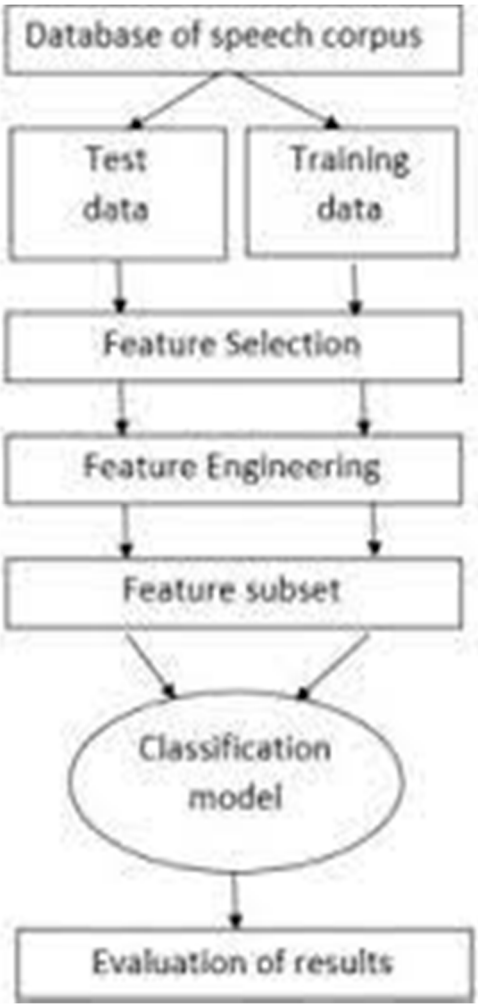
D. Feature Analysis

The process of altering, decreasing, or creating features for the dataset is known as feature engineering. Each feature has several values for each frame of the audio stream, as was originally noted in the raw data. The frame size and frame overlap values can be adjusted to obtain the correct values of the audio signal using the frame blocking and windowing methods. Also, the average values of various attributes for the audio signals are determined using the averaging technique. Each audio signal is now represented by 34 discrete values in the converted data. A critical choice to make is to reduce the number of features. Removal of features is typically dependent on subject-matter expertise, which can impact the system's performance.

III. MODELLING AND ANALYSIS

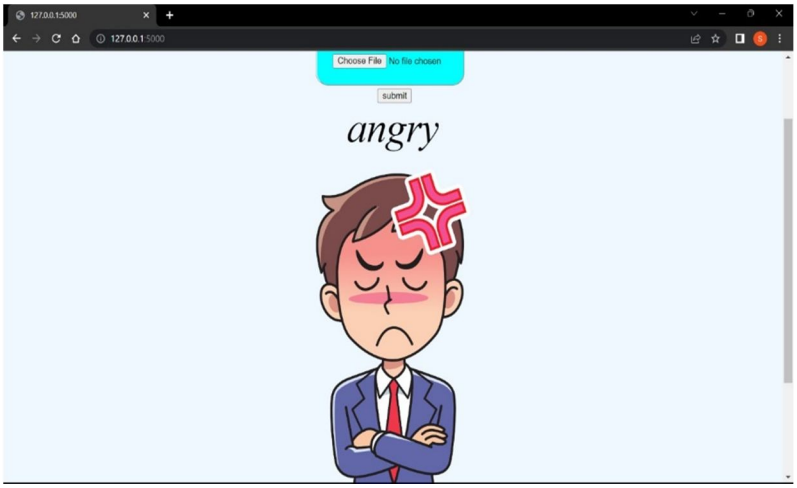
The top six emotions, or the "big n category of emotions," according to an analysis of 64 emotional speech data collections, are sadness, surprise, rage, joy, and fear. Other emotions are modeled into subcategories that are distinct from the basic categories, such as neutral, positive, and negative, as well as rage, joy, sadness, and fear, among other big n emotions. They are divided into variations on the primary categories, such as astonished, furious, and bored, etc. Another technique to model emotions is through dimension aspects, where the major consideration is how many dimensions are taken into account. "Arousal is a three-dimensional space to depict the states of human emotion. It is the individual's overall feeling of dynamism or lethargy. Activity includes both mental and physical preparation for action as well as overt activity. Power and control are similar ideas that fall under the power dimension, but the main concern with emotions is how people perceive their own power. An individual's overall sense of misery or well-being, which is tied to their circumstances, is reflected in this valence. Grammar changes or the words we choose play a significant effect in how our emotional state affects them. N-Grams and Bag-of-Words are two of the most popular analysis methods now in use. A probabilistic base language model and N-grams, a numerical representation of texts used in automatic document categorization, are utilized to forecast the next item in a given sequence. Before using this strategy, it is helpful to reduce speech complexity by removal and enhance a general minimum frequency of recurrence. Non-linguistic vocalizations like sobs sigh and cries can all be included in the language.

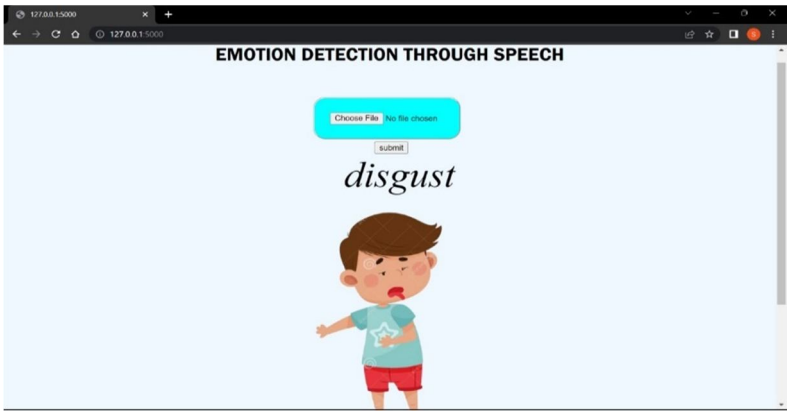
BLOCK DIAGRAM

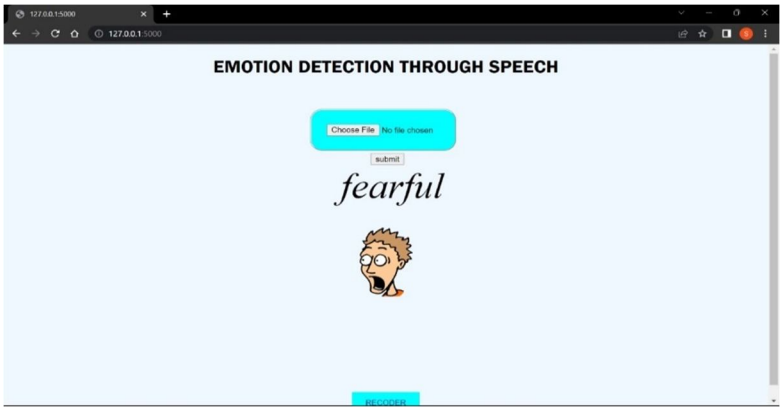


IV. RESULTS AND DISCUSSION

The results of the implementation can be used to draw a number of observations and inferences. The performance scores between the various techniques have improved generally. The implementation utilizing the first strategy performed ok with a high score of 83% using the SVM method, the second strategy worked well with a high score of 80% using the KNN algorithm, and the third strategy got the greatest score of 90% using the SVM algorithm.





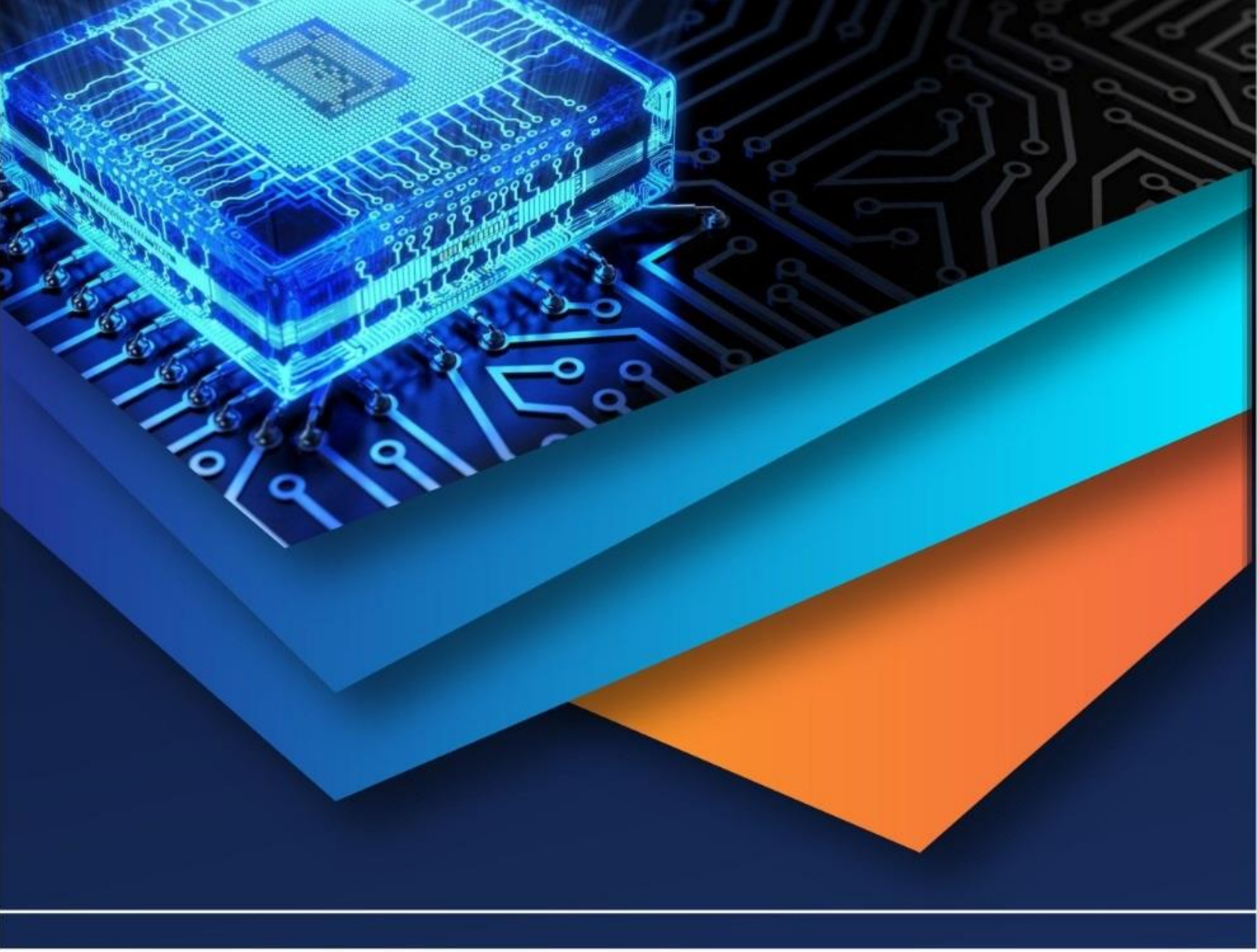


V. CONCLUSION

AI and machine learning's rapid growth and development have ushered in a new era of automation. The majority of these automated devices are controlled by the user's voice commands. Many advantages can be built over existing systems if machines can understand the speaker's emotion in addition to recognizing words (user). The applications of a speech emotion detection system include computer-based tutorial applications, automated call center conversations, a diagnostic tool for therapy, and an automatic translation system. In addition to emotions, the model can recognize feelings such as depression and mood changes. Therapists can use such systems to track their patients' mood swings. A sarcasm detection system is a difficult product of creating emotional machines. Sarcasm detection is a more difficult problem than emotion detection because sarcasm cannot be easily identified using only the speaker's words or tone. To identify possible sarcasm, sentiment detection using SPEECH EMOTION DETECTION 70 vocabulary can be combined with speech emotion detection.

REFERENCES

[1] M. Soegaard and R. Friis Dam (2013). The Second Edition of the Encyclopedia of Human-Computer Interaction
[2] Developer.amazon.com. (2018). (2018). Alexa from Amazon. [online] Alexa can be found at: https://developer.amazon.com/alexa
[3] Store.google.com. (2018). (2018). Google Home - Google Store. [online] Google Home Learn is available at https://store.google.com/product/google home to learn
[4] Apple. (2018). Siri is an iOS app. [online] You can find it at https://www.apple.com/ios/siri/.
[5] The Samsung Galaxy Official Website. (2018). What exactly is S Voice? [online] Website: http://www.samsung.com/global/galaxy/what-is/s-voice/ [Accessed May 2, 2018].
[6] Gartner.com. (2018). (2018). According to Gartner, 8.4 billion people are connected. [online] The following link is available: https://www.gartner.com/newsroom/id/3598917
[7] S. S. Narayanan, "Toward recognizing emotions in spoken dialogues," IEEE Trans. Speech Audio Process., vol. 13, no. 2, Mar. 2005, pp. 293–303.
[8] Pichora-Fuller and Dupuis, K. (2010). [Collection] Toronto emotional speech set, Psychology Department at the University of Toronto (TESS). Toronto.
[9] S. Garca, L. Sánchez, F. Herrera, J. Alcalá-Fdez, A. Fernández, J. Luengo, and J. Derrac. Data Set Repository for the KEEL Data gathering Software Tool



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY