**SPEECH EMOTION DETECTION USING MACHINE LEARNING**

**TECHNIQUESTECHNIQUES**

**TEAM - DATA SHIELDERS**

SURYA VAMSI PASUPULETI

RAM VENKAT BUDDALA

DHEERAJ

HARIKA

Contents

[Abstract 3](#_Toc132258381)

[Introduction 4](#_Toc132258382)

[METHODOLOGY 5](#_Toc132258383)

[DATASET 6](#_Toc132258384)

[TORONTO EMOTIONAL SPEECH SET (TESS) 6](#_Toc132258385)

[IMPLEMENTATION 6](#_Toc132258386)

[DATA COLLECTION 6](#_Toc132258387)

[PYTHON LIBRARY 7](#_Toc132258388)

[Results 8](#_Toc132258389)

[Fear 9](#_Toc132258390)

[Angry 10](#_Toc132258391)

[Disgust 11](#_Toc132258392)

[Neutral 12](#_Toc132258393)

[Sad 13](#_Toc132258394)

[PS 14](#_Toc132258395)

[Happy 15](#_Toc132258396)

[Accuracy 17](#_Toc132258397)

[Conclusion 20](#_Toc132258398)

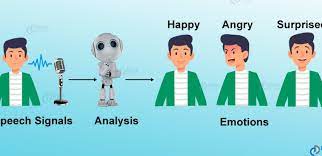
# Abstract

One must become skilled in the art of communication to be able to clearly explain their opinions and thoughts. Speech is by far and away the most prevalent and significant form of human communication. As the Internet of Things (IoT) age continues to evolve, access to intelligent systems is becoming progressively more readily available. This category contains a wide range of products, from simple items such as wearables and widgets to complex automated systems that are employed in a wide number of business sectors. Intelligent applications are ones that can be interacted with and utilized with minimum input from the user, most frequently through the use of voice. These apps are referred to as "hands-free" or "hands-free" apps. Because of this, it is necessary that these AI technologies completely comprehend the language that is uttered. A speech percept can provide information about the speaker's gender, age, language, and emotional state, all of which can be inferred from the percept. An analysis of the speaker's emotional state may be performed by integrating several currently available speech recognition algorithms with an emotion detection system. These approaches are employed in Internet of Things applications. The effectiveness of an emotion detection system has far-reaching ramifications for the utility and efficiency of applications made possible by the Internet of Things (IoT). This paper presents a voice emotion detection system that has advancements over an existing system in terms of the data, feature selection, and approach used. The goal of this study is to better recognize speech percepts based on emotions.

# Introduction

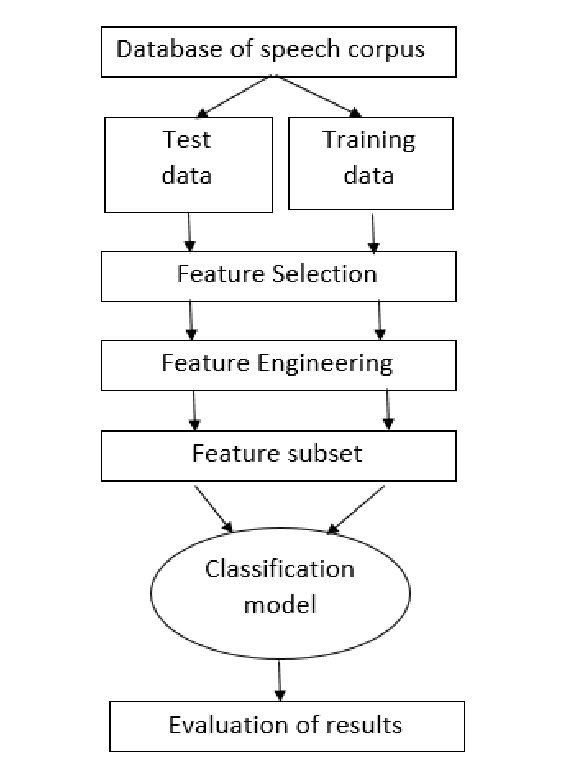
Speech emotion detection refers to the process of studying and analyzing the emotive aspects that are present in human speech. Emotional outbursts are an essential component of human contact as well as the dissemination of information. Automatic emotion identification from speech signals might potentially be of considerable use in a wide variety of contexts, including human-computer interaction, speech therapy, and psychological research, to name just a few of the many possible application areas. A recent surge in the field of machine learning has enabled researchers to develop more complex algorithms for analyzing the emotional state of a speaker based only on their speech patterns. Within the scope of this research, we investigate the landscape of machine learning techniques that may be used to the problem of determining the emotional tone of a speaker's voice. We provide an in-depth analysis of the advantages and disadvantages of each approach, as well as instances of its use in the real world. Our goal is to provide light on the existing landscape of speech emotion detection algorithms and the potential applications of these techniques [1].

Over the course of the last several years, advancements in artificial intelligence (AI) have been making remarkable strides. Artificial intelligence was once exclusively the purview of computer scientists, but it has since found its way into our everyday lives in the shape of various intelligent systems. The advancements that have been made in artificial intelligence have had a direct and positive impact on the growth of human-computer interface (HCI) technologies [1]. Because human-computer interaction (HCI) is the face of artificial intelligence (AI), which millions of people engage with, it is essential to concentrate on establishing and improving approaches for HCI. Some of the current methods of human-computer interaction (HCI) include the use of touch, movement, hand gestures, voice, and facial expressions. Intelligent devices that can be triggered by the user's voice are rapidly becoming the solution of choice in a variety of sectors. To accurately pick up the instructions that are sent to it, a computer agent working in a voice-based system has to have a complete understanding of the human's speech percept [2].



# METHODOLOGY

The construction of the speech emotion detection system involves the usage of an ML model. The implementation strategies are standard fare for any machine learning project; however, fine-tuning phases have been added to the process in order to optimize the performance of the model. A graphical depiction of the entire method may be found in the form of a flowchart. The initial step in the process is to collect all the pertinent information. All the assessments and results that are produced by a created model are informed by the data that the model has learnt from [3]. The information that was acquired is put through several different machine learning tasks in the second step, which is called feature engineering. These solutions resolve the issues with the quality of the data as well as the representation of the data. It is common practice to refer to the third phase of a machine learning project as the "meat and potatoes" phase since this is when an algorithm-based model is developed. This model makes use of an ML algorithm to analyses the data, adjust to any new information that comes its way, and learn from any mistakes it may have made in the past. After the construction of the model is finished, it must be evaluated to see whether or not it functions as designed. A cycle of model building and assessment is widely utilized by developers in the process of assessing how well various algorithms compare to one another. The results of these comparisons are helpful in selecting which machine learning approach is most appropriate for the task that must be accomplished at hand.



# DATASET

## TORONTO EMOTIONAL SPEECH SET (TESS)

A speech emotion-based dataset was developed in 2010 in the English language by academics from the Department of Psychology at the University of Toronto. There are 2800 audio tracks in the database, and they represent seven different fundamental emotions: happy, sad, angry, surprised, afraid, disgusted, and neutral. It's a performance tape with voices from two different generations reading the script: an older gentleman (age 64) and a younger man (age 26) [4].

The following are some of the characteristics of the dataset that make it suitable for this work:

* To train an efficient model, a dataset of sufficient size is required. A model's effectiveness increases as it is exposed to additional data. All the essential types of data for modelling emotions are there. Additional studies, such as those focusing on the detection of sarcasm and depression, can benefit from a mix of these feelings [5].
* The categorization is enhanced by collecting data from two distinct age groups.
* The audio files are mono signals; thus, they may be converted without any problems using the vast majority of existing software libraries.

# IMPLEMENTATION

## DATA COLLECTION

The initial step in the process of developing a Speech Emotion Recognition system is collecting audio samples that have been annotated with a variety of emotions for the purpose of training the model. You may find the audio snippets on the internet in the form of downloadable wav or mp3 files. In this section, we will discuss the processes that were utilized in combination with the investigations that were carried out on the TESS dataset.

## PYTHON LIBRARY

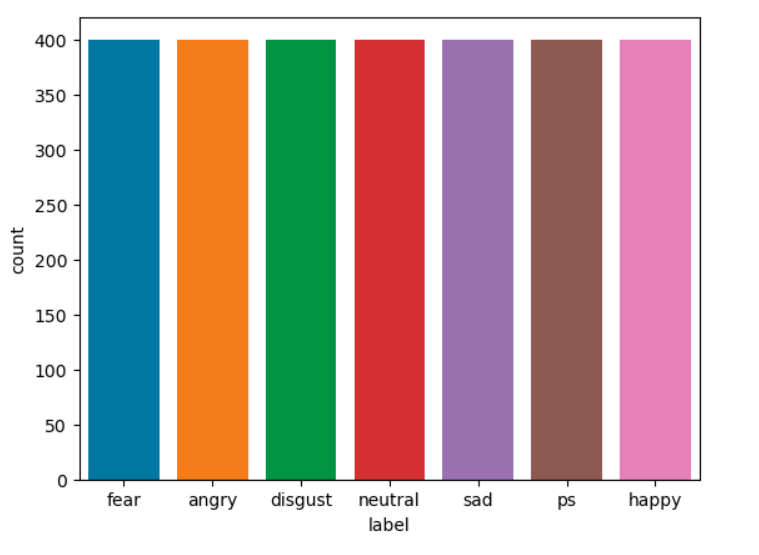
Following the completion of the initial step of collecting data, the following phase consisted of translating the audio recordings into a numerical representation so that additional analysis could be carried out. The process of obtaining numerical values for different characteristics of an audio file is referred to as feature extraction. The pyAudio Analysis package was utilized in order to accomplish this goal. This Python package contains methods for the extraction of features, including a windowing system that allows for customizable parameters such as frame size and frame step. Following the completion of this procedure, a comma-separated values (CSV) file was produced. This file included 34 columns that each represented a different characteristic, and it had one row for each audio file. The number of frames in an audio file will be used to establish the range of possible values for each attribute [6]. The pyAudio Analysis library is an open-source Python library that specializes in finding solutions to issues pertaining to the extraction of features, categorization of data, segmentation of data, and visualization of data in the field of audio. The functionality of the library is dependent on the presence of several additional libraries:

* NumPy
* Matplotlib
* SciPy
* Sklearn
* Hmmlearn
* Simplejson
* eyeD3
* pydub

# Results

After reading in an emotional audio dataset, the algorithm proceeds to extract Mel-frequency cepstral coefficients (MFCCs) from the audio files by using the librosa package. This is done after reading in an emotional audio dataset. These qualities serve as input for the construction of a one-hot encoded representation of the emotions that are contained within the dataset. After that, we design a sequential model with the help of the Keras package. This model has LSTM layers, dropout layers, and dense layers. After that, the model is constructed using the optimizer of Adam, the measure of accuracy, and the loss function of categorical cross entropy.

After then, the model is trained on the input data for a total of fifty epochs, using a batch size of sixty-four and a validation split of twenty-five percent. At long last, the loss and accuracy during the training and validation process are graphed using matplotlib. In conclusion, it looks as though the code provides a framework for constructing an MFCC- and LSTM-based machine learning model that is capable of identifying emotions conveyed through spoken language. Without hard data on the model's performance, it is difficult to determine whether it is effective.

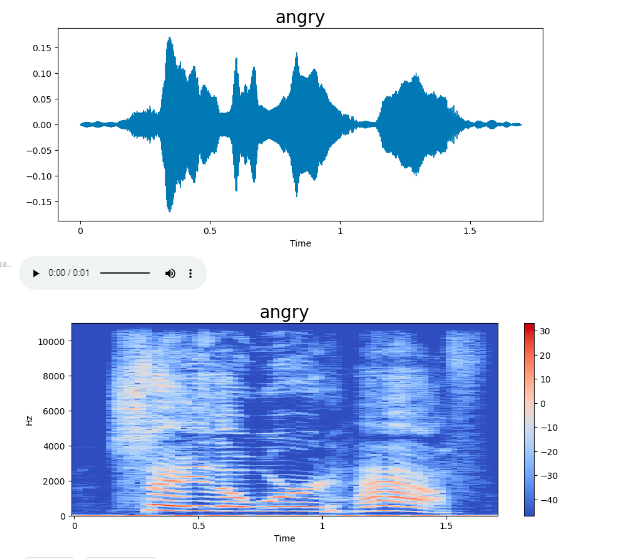


## Fear

Diagram

Description automatically generated

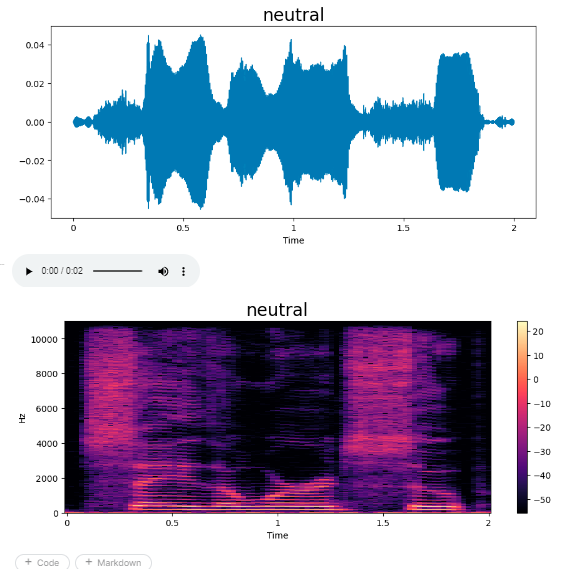
## Angry



## Chart, scatter chart Description automatically generated with medium confidenceDisgust



## Neutral





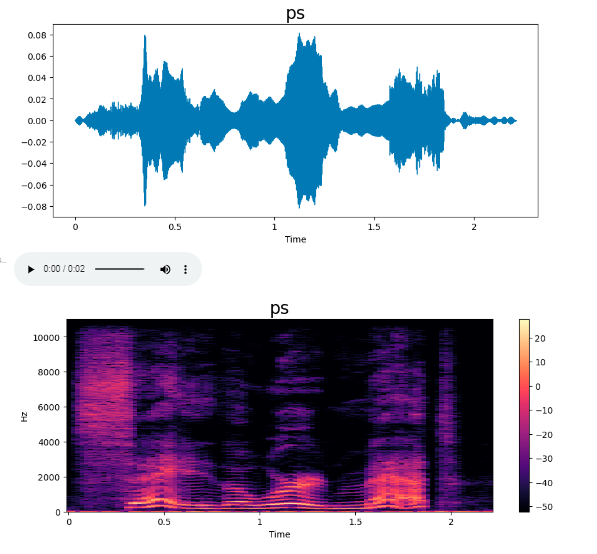
## Sad

Chart

Description automatically generated



## PS





## Happy

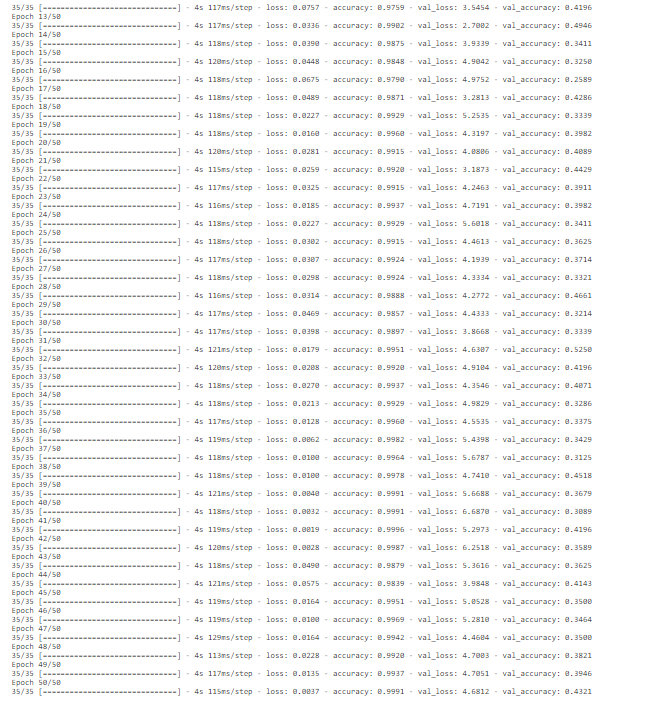


Graphical user interface

Description automatically generatedGraphical user interface

Description automatically generated

## Accuracy



Chart

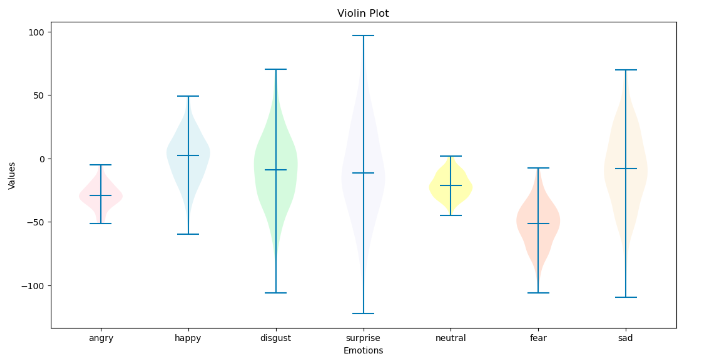
Description automatically generatedChart, histogram

Description automatically generated

Distribution of the data on each of its feature can be visualized using box plots and violin plots or using  
pair plots. The Seaborn package in Python provides methods to draw such plots. Shown here is the data  
distribution of the feature ‘Energy’ among the different categories

Chart, box and whisker chart

Description automatically generated



# Conclusion

The latest advancements in artificial intelligence and machine learning have ushered in a new era of automation. The vast majority of these automated solutions have the capability to react to the user's voice commands. In addition to identifying the words, robots that could comprehend the disposition of the person speaking (the user) would allow for a number of potential enhancements to be made to the existing systems. A vocal emotion detection system can have various applications, some of which include automatic translation, a diagnostic tool for treatment, and computer-based educational applications. These are just a few examples.

In this thesis, the steps that are involved in constructing a speech emotion recognition system are laid out, and experiments are undertaken to identify the consequences of various choices. Ultimately, a speech emotion recognition system is developed. Due to the limited amount of publicly accessible speech datasets, it was initially challenging to deploy a model that had been adequately trained. After then, several distinct approaches to feature extraction were provided in the earlier papers, and a great deal of testing was necessary to determine which approach was the most effective. Finally, in order to pick a classifier, it was necessary to conduct research on the relative merits of a number of different classifiers and how effectively they functioned in determining emotional states. The findings of the trials make it abundantly evident that making use of a multitude of traits in conjunction with one another leads in a better recognition rate than making use of only one.

It is possible to model the proposed project in order to gain insight into how to make changes in the future about its usefulness, accuracy, and efficacy. There is a possibility that the model will experience negative emotions such as hopelessness and mood swings. This sort of strategy might be utilized by therapists in order to monitor the patients' emotional states more effectively. Building emotionally aware computers comes with a number of challenging side effects, one of which is the creation of a system that can recognize sarcasm. Sarcasm identification is a more challenging problem than emotion recognition since it cannot be easily established based just on the words or tone of the speaker. Analysis of sentiment performed by language, can be used in conjunction with speech emotion detection to determine if sarcasm is present. As a result, there will be several uses for a speech-based emotion identification system in the near future.

# References

1. Khalil, R. A., Jones, E., Babar, M. I., Jan, T., Zafar, M. H., & Alhussain, T. (2019). Speech emotion recognition using deep learning techniques: A review. *IEEE Access*, *7*, 117327-117345.
2. El Ayadi, M., Kamel, M. S., & Karray, F. (2011). Survey on speech emotion recognition: Features, classification schemes, and databases. *Pattern recognition*, *44*(3), 572-587.
3. Ingale, A. B., & Chaudhari, D. S. (2012). Speech emotion recognition. *International Journal of Soft Computing and Engineering (IJSCE)*, *2*(1), 235-238.
4. Toronto emotional speech set (TESS) | TSpace Repository. (n.d.). https://tspace.library.utoronto.ca/handle/1807/24487
5. Swain, M., Routray, A., & Kabisatpathy, P. (2018). Databases, features and classifiers for speech emotion recognition: a review. *International Journal of Speech Technology*, *21*, 93-120.
6. Li, R., Liang, Y., Liu, X., Wang, B., Huang, W., Cai, Z., ... & Pan, J. (2021). MindLink-eumpy: an open-source python toolbox for multimodal emotion recognition. *Frontiers in Human Neuroscience*, *15*, 621493.