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.2023.49703

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 Recognition Using Deep Learning

Prof. Jagdish Kambale[1], Abhijeet Khedkar[2], Prasad Patil[3], Tejas Sonone[4]
[1, 2, 3, 4]*Information Technology Department, Pune Institute of Computer Technology, Pune*

Abstract: ***Due to different technical developments, speech signals have evolved into a kind of human-machine communication in the digital age. Recognizing the emotions of the person behind his or her speech is a crucial part of Human-Computer Interaction (HCI). Many methods, including numerous well-known speech analysis and classification algorithms, have been employed to extract emotions from signals in the literature on voice emotion recognition (SER). Speech Emotion Recognition (SER) approaches have become obsolete as the Deep Learning concept has come into play. In this paper, the algorithm for identifying speech-based emotions is implemented using deep learning. It also provides an overview of deep learning methodologies and examines some recent research that makes use of these methods. It makes use of a dataset of various emotional voices and then aids in the identification of that emotion. It will be beneficial for the computers or robots to understand humans more clearly and function in accordance with it.***

Keywords: Deep Learning, Speech Processing, Human Computer Interaction (HCI), Python, Data Science, Dataset, Neural Network, Long Short-Term Memory (LSTM), pandas

I. INTRODUCTION

A. Introduction

Speech Emotion Recognition (SER) is the process of identifying emotional components in speech regardless of semantic content. Although this is a natural part of human speech communication, it is still being researched whether it can be performed automatically by programmable devices. The curiosity with voice emotion recognition was the primary factor in choosing this problem statement. Deep learning as a method of speech emotion recognition seems to be a highly creative notion. The subject of artificial intelligence has a very broad scope. Enhancing voice emotion recognition will significantly support augmented reality, machine learning, virtual reality, and human computer interface. The purpose of research on automatic emotion detection systems is to provide effective, real-time algorithms for identifying emotions in a wide range of human-machine interaction users, such as call centre workers and customers, drivers, pilots, and others. Emotions have been shown to be necessary for making machines appear and behave more humanlike. Emotionally intelligent robots may behave correctly and demonstrate. People with emotive personalities In some circumstances, computer-generated characters capable of speaking in a highly realistic and convincing manner by appealing to human emotions could stand in for individuals. Speech-based emotional expressions must be interpreted by machines. This capacity is required to have a completely meaningful discussion based on mutual human-machine trust and comprehension.

B. Scope and Objectives

The speech emotion recognition will be able to provide a platform for : Wide range of application in Internet of things. Communication and intelligent analysis between virtual objects and humans in Metaverse. Improve interaction Between Human and computer in later years gradually. The following are the project's objectives:

Develop a Deep Learning Model to analyse and classify Speech Emotion of various types. Use Algorithms to describe the real time working of model to end user. Develop an Application Interface to communicate the interpretations to the end user By merging several datasets and producing a comparative reported create an accurate model.

II. LITERATURE REVIEW

While reading through multiple different papers and websites we found that there were a few common problems that the people were facing like the use of Internet of Things in emotion recognition so predicting and monitoring all has become tough. Another problem is that less interaction between human and computer.

Speaker-sensitive emotion recognition via ranking presents a ranking approach for emotion recognition that integrates information about speaker's overall expressivity. [1] 'H. Cao, R. Verma, and A. Nenkova' proposes that their strategy leads to significant improvements in accuracy when compared to standard approaches.

©IJRASET: All Rights are Reserved

Speech emotion recognition: Features and classification models [2] 'L. Chen, X. Mao, Y. Xue, and L. L. Cheng' proposes a three-level speech emotion recognition model which is used to classify six speech emotions, including sadness, anger, surprise, fear, happiness, and disgust, from coarse to fine, in order to address the speaker independent emotion recognition problem. Fisher rate, which is also considered an input parameter for Support Vector Machine, is used to choose acceptable features from 288 possibilities for each level (SVM). . The experimental results demonstrated that Fisher outperforms PCA for dimension reduction and SVM outperforms ANN for speaker independent speech emotion recognition.

Automatic speech emotion recognition using modulation spectral features [3] 'S. Wu, T. H. Falk, and W.-Y. Chan' puts Modulation spectral features (MSFs) in this paper for the automatic detection of human emotive information from speech. The features are derived from a long-term spectro-temporal representation. In a test of categorization of discrete emotion categories, the MSFs outperform features based on mel-frequency cepstral coefficients and perceptual linear prediction coefficients, two extensively used short-term spectrum representations. motion Recognition of Affective Speech Based on Multiple Classifiers Using Acoustic-Prosodic Information and Semantic Labels [4] 'C.-H. Wu and W.-B. Liang' describes a method for recognising affective speech emotions utilising several classifiers and acoustic-prosodic information (AP). Acoustic and prosodic information such as spectrum, formant, and pitch-related features are retrieved from the detected emotional salient parts of the input speech for AP-based recognition. To acquire the AP-based emotion recognition confidence, a Meta Decision Tree (MDT) is used for classifier fusion. Finally, for the ultimate emotion judgement, a weighted product fusion method is employed to fuse the AP-based and SL-based recognition results.

III. SYSTEM REQUIREMENTS

A. *Functional Requirements*

1) *Dataset:* Training Deep learning models requires extensive data to achieve high accuracy, low loss of features and increasing efficiency. The project uses Torronto Emotion Speech Set(TESS).
2) *Data Pre-processing:* Training Deep Learning models on large datasets requires effective data pre-processing under consideration for effective feature extraction. We have use various data Pre-processing techniques for same.
3) *Emotion Recognition capability:* After Pre-processing, a Deep learning model is trained for effective human recognition to generate accurate results.
4) *Audio Pre-processing capability:* for real time audio detection, effective audio processing capability is needed to be able to detect and recognize the data conveyed through audio.
5) *Model Interpretability:* for the effective recognition, effective models like LSTM are used ti explain the working of Models used.

B. *Non-Functional Requirements*

1) *Testability:* for effective working of model, modular architecture is used where each part is divided into multiple modules for correct working.
2) *Reliability:* Feature tests are performed to ensure reliability of dataset and quality of dataset, So the accuracy and performance of system are satisfactory.

IV. PROPOSED METHODOLOGY

To tackle the problem of human interaction with computers with voice synchronization and differentiation we proposed following methodology: Develop a Deep Learning Model to analyse and classify Speech Emotion of various types. Process the available datasets of wide variety to test data and generate defined output. Use LSTM model, MFCC coeficient for effective expansion of problem statement. Use Machine Learning libraries to visualise the audio files and analysis of results.

A. *Data Pre-processing*

Training For deep learning models to attain high accuracy, less feature loss, and growing efficiency, they need a lot of data. The task makes use of the Toronto Emotion Speech Set (TESS). Effective data pre-processing is necessary for effective feature extraction when training Deep Learning models on huge datasets. For the same, we have used a variety of data preprocessing approaches.

B. *Emotion Recognition capability*

Information extraction is the first step in the process of generating an audio feature that is used to automatically extract knowledge units from audio information from various sources. After pre-processing, a Deep Learning model is trained for effective human recognition to generate trustworthy results.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

C. Audio Prep-processing capability

A topological map of the audio sources as a function of time delay is produced by processing the audio waveforms. Animals can localise audio sources effectively utilising a number of auditory clue. for real time audio detection, effective audio processing capability is needed to be able to detect and recognize the data conveyed through audio.

D. Model Interpretability

When people can grasp the justification for the predictions and judgments made by the model, it is said that the model is interpretable. It is simpler for someone to understand and trust a model the more interpretable it is.

V. SYSTEM ARCHITECTURE

The Fig.1 shows the complete system architecture of the Speech emotion recognition. This includes all the proposed methodology that is the signal acquisition, Feature Extraction, Training-Testing ,Emotion Classification The system consists of Web interface, A speech Emotion recognition Module and a Integration Module with camera with the interface for real time emotion recognition. Emotion Recognition Module Working is given in the diagram.

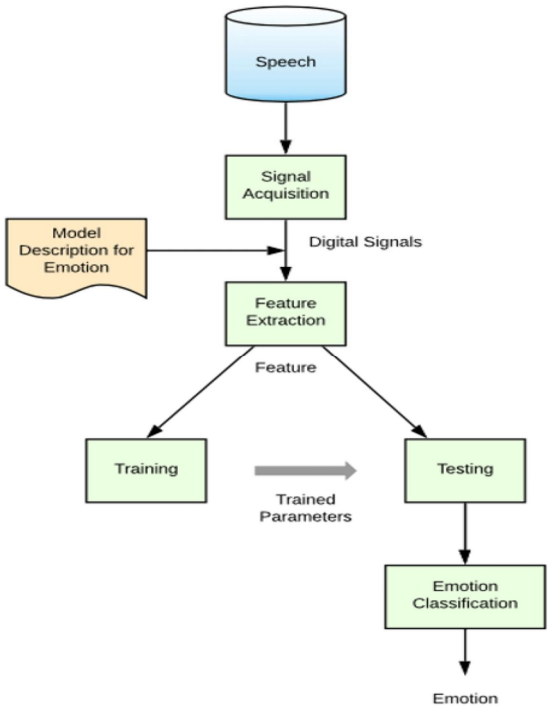


Figure 1 : System Architecture

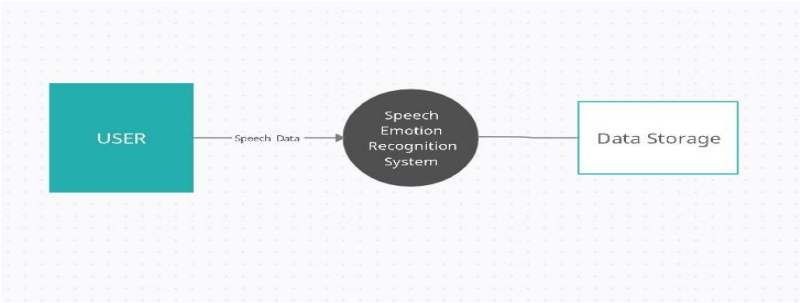


figure 2: Data Flow Diagram

VI. METHODOLOGIES

There are set of 200 target words were spoken in the carrier phrase ”Say the word” and recordings were made of the set portraying each of seven emotions (anger, disgust, fear, happiness, pleasant surprise, sadness, and neutral). There are 2800 data points (audio files) in total.Then, using existing data to create new data points, we enhance the original dataset to artificially increase the amount of data. Following that, we develop a model that is trained using the data that was gathered; this model is essentially a Recurrent neural network in which the input audio file passes through a number of filters. We then extract Acoustic Features from the given input audio though MFCC and thus get mathematical parameters using which we get a collection of pints upon which classification is to be made.

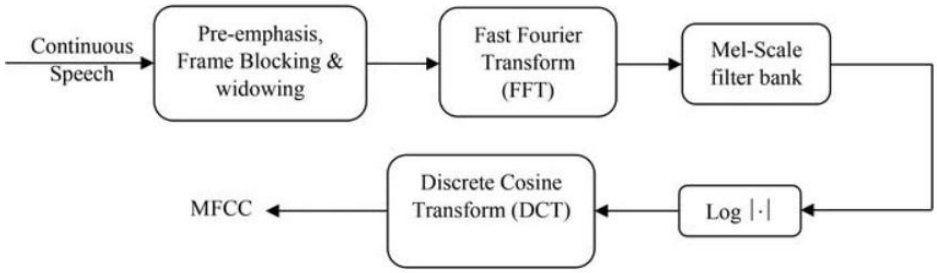


figure 3: MFCC processor

VII. CONCLUSION AND FUTURE SCOPE

A. *Conclusion*

In this research, we tried to use deep learning to analyse a few voice samples. We first loaded the datasets, and then, using the Librosa library’s functions waveshow and spectrogram, we depicted the various human emotions. Then we constructed the LSTM model, trained it, and then used the Matplotlib tool to depict the data in graphical form. After repeated testing with various settings, the model’s average accuracy was discovered to be 71 percent..

B. *Future Scope*

As a result, speech emotion recognition is a fascinating topic about which there is still much to discover. Future work on our model will concentrate on improving the model’s accuracy to get better outcomes. Additionally, the model can be trained to produce results for speeches that are longer than those employed in this model. Additionally, I’ll create a website to showcase the results of my future efforts. Additionally, employ the Speech Emotion Recognition Model.

VIII. ACKNOWLEDGMENT

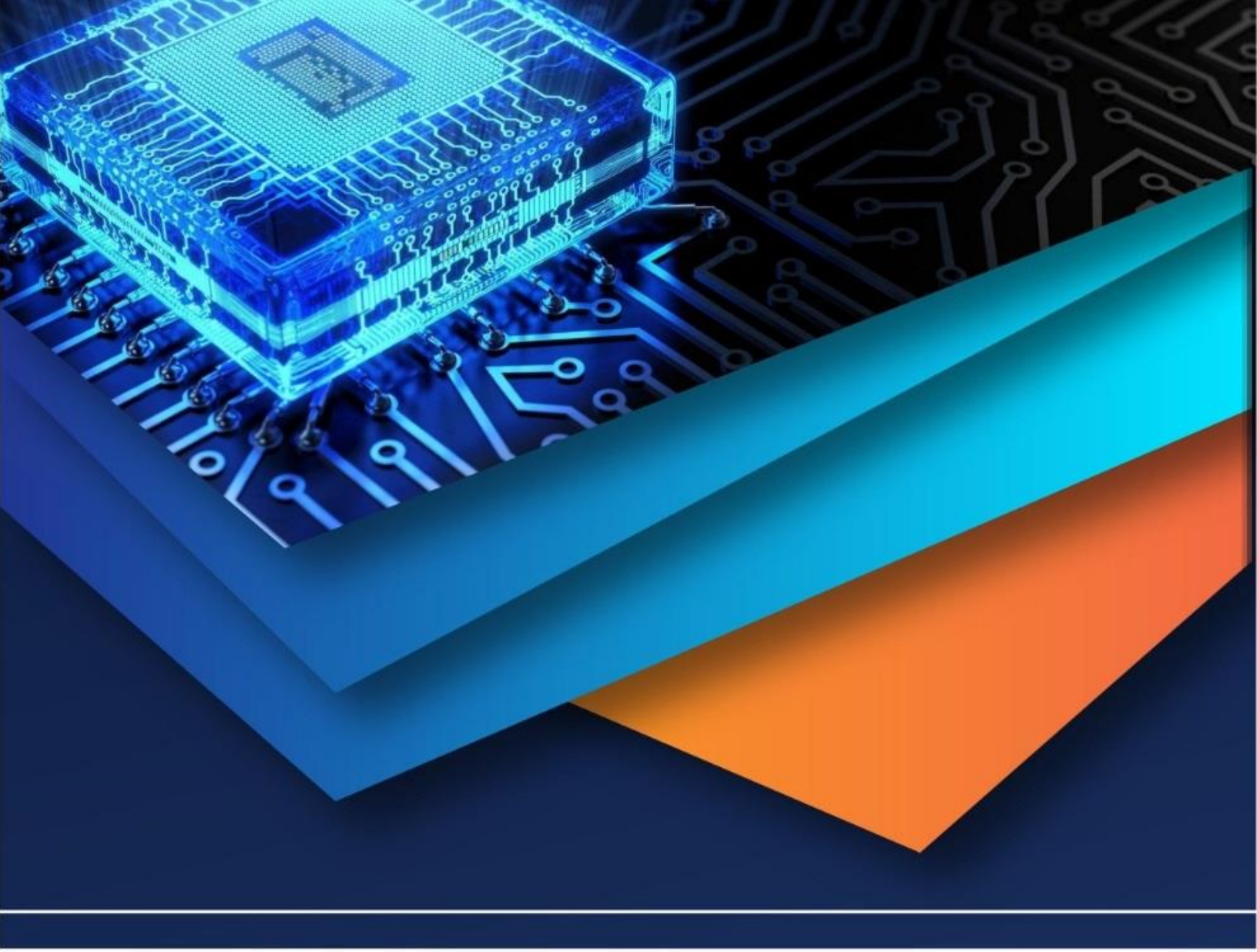
The satisfaction that accompanies the successful completion and it asks would be incomplete without the mention of the people who made it possible and whose constant encouragement and guidance have been a source of inspiration throughout the course of this project. We take this opportunity to express my sincere thanks to my guide respected Prof. Jagdish K. Kambale sir for their support and encouragement throughout the completion of this project. Finally, We would like to thank all the teaching and non-teaching faculty members and lab staff of the department of Information Technology for their encouragement. We also extend our thanks to all those who helped us directly or indirectly in the completion of this project. We are also thankful to our reviewers Prof. Abhinay Dhamankar and Prof.Radhika Kulkarni for their valuable suggestions.We are grateful to Dr.Archana S. Ghotkar, Head of Department of Information Technology, Pune Institute of Computer Technology for her indispensable support and suggestions..

REFERENCES

[1] 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.
[2] 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
[3] S. Wu, T. H. Falk, and W.-Y. Chan, “Automatic speech emotion recognition using modulation spectral features,” Speech Commun., vol. 53, no. 5, pp. 768–785, May 2011.
[4] C.-H. Wu and W.-B. Liang, “Emotion Recognition of Affective Speech Based on Multiple Classifiers Using Acoustic-Prosodic Information and Semantic Labels,” IEEE Trans. Affect. Comput., vol. 2, no. 1, pp. 10–21, Jan. 2011

[5] E. M. Albornoz, D. H. Milone, and H. L. Rufiner, "Spoken emotion recognition using hierarchical classifiers," Comput. Speech Lang., vol. 25, no. 3, pp. 556–570, Jul. 2011.
[6] Wu et al. proposed a fusion-based method for speech emotion recognition by employing multiple classifier and acoustic-prosodic (AP) features and semantic labels (SLs).
[7] Narayanan proposed domain-specific emotion recognition by utilizing speech signals from call center application. Detecting negative and nonnegative emotion (e.g. anger and happy) are the main focus of this research.
[8] Yang & Lugger presented a novel set of harmony features for speech emotion recognition. These features are relying on psychoacoustic perception from music theory.
[9] B. W. Schuller, ''Speech emotion recognition: Two decades in a nut shell, benchmarks, and ongoing trends,'' Commun. ACM, vol. 61, no. 5, pp. 90– 99, 2018.
[10] N. D. Lane and P. Georgiev, ''Can deep learning revolutionize mobile sensing?'' in Proc. ACM 16th Int. Workshop Mobile Comput. Syst. Appl., 2015, pp. 117–122.
[11] A. B. Nassif, I. Shahin, I. Attili, M. Azzeh, and K. Shaalan, ''Speech recognition using deep neural networks: A systematic review,'' IEEE Access, vol. 7, pp. 19143–19165, 2019.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)