IJNRD.ORG ISSN : 2456-4184
INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD) | IJNRD.ORG
An International Open Access, Peer-reviewed, Refereed Journal

Speech Emotion Recognition using Machine Learning in Python

GOPIKRISHNAN S
Student
MADHA ENGINEERING COLLEGE

CHAPTER 1

INTRODUCTION

For several years now, the growth in the field of Artificial Intelligence (AI) has been accelerated. AI, which was once a subject understood by computer scientists only, has now reached the house of a common man in the form of intelligent systems. The advancements of AI have engendered 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 user's experiences. Some of the existing HCI methods involve communication through touch, movement, hand gestures, voice and facial gestures [1]. Among the different methods, voice-based intelligent devices are gaining popularity in a wide range of applications. In a voice- based system, a computer agent is required to completely comprehend the human's speech percept in order to accurately pick up the commands given to it. This field of study is termed as Speech Processing and consists of three components:

- Speaker Identification
- Speech Recognition
- Speech Emotion Detection

Speech Emotion Detection is challenging to implement among the other components due to its complexity. Furthermore, the definition of an intelligent computer system requires the system to mimic human behavior. A striking nature unique to humans is the ability to alter conversations based on the emotional state of the speaker and the listener. Speech emotion detection can be built as a classification problem solved using several machine learning algorithms. This project discusses in detail the various methods and experiments carried out as part of implementing a Speech Emotion Detection system.

1.1 IMPORTANCE

Communication is the key to expressing oneself. Humans use most of their body and voice to effectively communicate. Hand gestures, body language, tone and temperament are all collectively used to express one's feelings. Though the verbal part of communication varies by languages practiced across the globe, the non-verbal part of communication is the expression of feeling which is most likely common among all. Therefore, any

advanced technology developed to produce a social environment experience also covers understanding emotional context in speech. Improvements in the field of emotion detection positively impact on a multitude of applications. Some of the research areas that benefit from automating the emotion detection technique include psychology, psychiatry, and neuroscience. These departments of cognitive sciences rely on human interaction, where the subject of studyis put through a series of questions and situations, and based on their reactions and responses, several inferences are made. A potential drawback occurs as few people are classified as introverts and hesitate to communicate. Therefore, replacing the traditional procedures with a computer-based detection system can benefit the study. Similarly, the practical applications of speech-based emotion detection are many. Smart home appliances and assistants (Examples: Amazon Alexa [2] and Google Home [3]) are ubiquitous these days. Additionally, customer care-based call centers often have automated voice control which mightnot please most of their angry customers. Redirecting such calls to a human attendant will improve the service. Other applications include eLearning, online tutoring, investigation, personal assistant (Example: Apple Siri [4] and Samsung S Voice [5]) etc. A very recent application could be seen in self-driving cars. These vehicles heavily depend on voice-based controlling. An unlikely situation, such as anxiety, can cause the passenger to utter unclear sentences. In these situations, understanding the emotional content expressed becomes of prime importance.

CHAPTER 2

LITERATURE SURVEY:

TITLE: Speech Emotion Recognition using Machine Learning

AUTHOR: Saroja R and Sreelekha G,2021

DESCRIPTION:

"Speech Emotion Recognition Using Deep Neural Networks and Transfer Learning" by Saroja R and Sreelekha G, published in IEEE Access in 2021. This paper proposed a novel approach for SER using deep neural networks and transfer learning techniques.

"Speech Emotion Recognition Using Convolutional Neural Networks with Transfer Learning" by Huan Wang et al., published in IEEE Access in 2020. This study proposed a SER system based on convolutional neural networks with transfer learning, achieving high recognition accuracy.

"Speech Emotion Recognition Using Hybrid Models of Deep Learning and Machine Learning Algorithms" by Jitendra Singh and Sunita S. Nair, published in the International Journal of Speech Technology in 2019. This paper proposed a hybrid approach for SER using deep learning and machine learning algorithms, achieving high accuracy on the IEMOCAP dataset.

"A Comprehensive Survey on Speech Emotion Recognition" by Arashdeep Kaur and Jaspreet Kaur, published in the Journal of Ambient Intelligence and Humanized Computing in 2020. This paper provided a comprehensive review of the state-of-the-art in SER, covering the different approaches and techniques used.

"Speech Emotion Recognition Based on Long Short-Term Memory and Random Forest" by Yi Chen et al., published in the Journal of Ambient Intelligence and Humanized Computing in 2020. This study proposed a SER system based on long short-term memory and random forest classifiers, achieving high recognition accuracy.

These studies demonstrate the diverse range of approaches and techniques used for SER using machine learning, highlighting the potential for further development and improvement in this field.Machine learning algorithms can be trained to recognize emotions in multiple languages.

Chapter 3

3. Problem Definition and Methodology

3.1 METHODOLOGY

The speech emotion detection system is implemented as a Machine Learning (ML) model. The steps of implementation are comparable to any other ML project, with additional fine-tuning procedures to make the model function better. The flowchart represents a pictorial overview of the process (see Figure 1). The first step is data collection, which is of prime importance. The model being developed will learn from the data provided to it and all the decisions and results that a developed model will produce are guided by the data.

The second step, called feature engineering, is a collection of several machine learning tasks that are executed over the collected data. These procedures address several data representation and data quality issues. The third step is often considered the core of an ML project where an algorithmic based model is developed.

This model uses an ML algorithm to learn about the data and train itself to respond to any new data it is exposed to. The final step is to evaluate the functioning of the built model. Very often, developers repeat the steps of developing a model and evaluating it to compare the performance of different algorithms. Comparison results help to choose the appropriate ML algorithm most relevant to the problem.

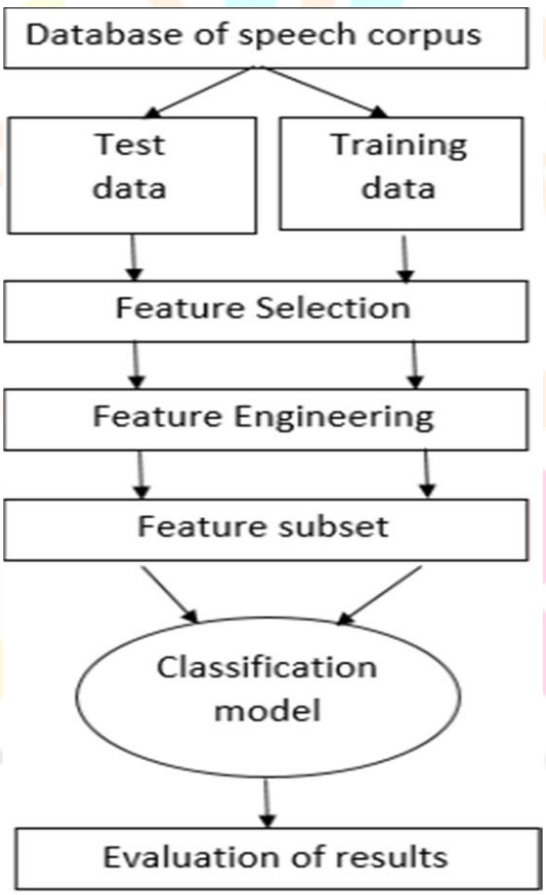


Fig.1 Flow of implementation

3.2 Scope

ML models used for SER are typically trained on large datasets of audio recordings that are label with emotional categories such as happy, sad, angry, and neutral. These models can then be used to automatically recognize emotions in new audio recordings.

The development of SER models using ML involves several stages, including feature extraction, model training, and model evaluation. Feature extraction involves extracting relevant features from audio signals, such as pitch, energy, and spectral characteristics. The extracted features are then used to train ML models such as Support Vector Machines (SVMs), Decision Trees, Random Forests, and Deep Neural Networks (DNNs).

Model evaluation involves testing the trained model on a separate dataset to evaluate its performance in recognizing emotions. Performance metrics such as accuracy, precision, recall, and F1 score are used to measure the performance of the model.

SER using ML has numerous applications in areas such as human- computer interaction, sentiment analysis, and affective computing. It can be used to develop emotion recognition systems for improving customer service, designing virtual assistants that can interact with users in a more natural way, and developing emotion-aware educational technologies.

CHAPTER 4

SYSTEM ANALYSIS

EXISTING SYSTEM

OpenSMILE is an open-source tool for extracting acoustic features from speech signals. It includes a set of pre-defined feature extraction configurations that can be used for SER.

EmoReact is a deep learning-based system that recognizes emotions from speech using a combination of spectral features and prosodic features.

PyAudioAnalysis is a Python library for audio analysis that includes a module for SER. It uses a combination of feature extraction techniques and machine learning algorithms, such as support vector machines (SVMs) and random forests.

4.1 PROPOSED SYSTEM

Dataset: The first step in building an SER system is to collect a large dataset of speech samples that cover a wide range of emotions. This dataset will be used to train and evaluate the machine learning models.

Feature Extraction: The next step is to extract relevant features from thespeech signal that can help in recognizing the emotions. Popular features used in SER include Mel-frequency cepstral coefficients (MFCCs), prosodic features, and spectral features.

Machine Learning Model: A machine learning model is then trained on the extracted features to classify emotions in speech signals. Various machine learning models can be used for SER, including Support Vector Machines (SVMs), Random Forests, and Neural Networks.

Model Evaluation: The trained model is evaluated on a test dataset to measure its performance. Various metrics can be used to evaluate the model, such as accuracy, precision, recall, and F1- score.

Deployment: Once the model is trained and evaluated, it can be deployed in real-world applications to recognize emotions from speech signals.

4.2 Requirement Specification

Functional Requirement

Speech Input: The system should be able to accept speech inputs in various formats, such as audio files or real-time audio streams.

Preprocessing: The system should preprocess the speech signal to remove noise and other unwanted components that may interfere with emotion recognition.

Feature Extraction: The system should extract relevant features from the preprocessed speech signal that can be used for emotion recognition, such as MFCCs, prosodic features, and spectral features.

Machine Learning Model: The system should have a machine learning model that can be trained on the extracted features to recognize emotions in speech signals.

Training and Evaluation:

The system should be able to train the machine learning model on a large dataset of speech samples and evaluate its performance on a separate test dataset.

Emotion Classification:

The system should be able to classify emotions in speech signals based on the trained machine learning model, such as happiness, sadness, anger, fear, and neutral.

Non-Functional Requirement:

Performance:

The system should be able to process speech signals in real-time or near real-time to provide timely and accurate results. It should also be able to handle a large volume of speech data and perform efficiently on different hardware configurations. Accuracy:

The system should have a high level of accuracy in recognizing emotions from speech signals. The accuracy should be measured using appropriate metrics such as F1-score or AUC-ROC.

Robustness:

The system should be able to perform well on different types of speech signals, including different accents, languages, and background noise levels.

Security:

The system should ensure the privacy and confidentiality of the speech data collected from users. It should also be protected against attacks such as data tampering, unauthorized access, and denial of service.

Usability: The system should have a user-friendly interface that is easy touse and understand. The system should also provide appropriate feedback to users to help them understand the results and improve their experience.

Scalability: The system should be designed to handle an increasing number of users and data volume over time. It should also be able to adapt to changing business requirements.

Maintainability: The system should be easy to maintain and update overtime. The code should be well-documented, modular, and adherent to good coding practices to ensure ease of maintainability.

4.5 Feasibility Study

4.5.1 Technical Feasibility

Machine Learning algorithms have been shown to be effective in SER, and there are many approaches that have been proposed and evaluated.One common approach is to use a combination of feature extraction techniques and supervised learning algorithms to recognize emotions from speech signals. These algorithms can be trained on labeled datasets of speech samples that are annotated with emotion labels.

Feature extraction techniques involve extracting relevant information from speech signals, such as pitch, energy, and spectral features, which can then be used to train machine learning models. There are also deep learning approaches, such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), that can learn to automatically extract relevant features from speech signals.

4.5.2 Economical Feasibility

From an economical feasibility perspective, the potential benefits of implementing SER using ML are numerous. One of the most significant benefits is the potential for cost savings in various industries. For example, in the healthcare industry, SER can be used to monitor the emotional state of patients in real-time, which can help healthcare professionals to intervene quickly and provide more targeted treatment. This can lead to faster recovery times and lower healthcare costs in the long run.

In the education sector, SER can be used to evaluate the emotional state of students, which can help teachers to adjust their teaching methods to better suit the needs of their students. This can lead to better academic performance, higher retention rates, and a more positive learning experience for students. In addition, SER can be used in the entertainment industry to create more engaging and immersive experiences for users.

4.5.3 Operational feasibility

Availability of data: One of the main requirements for building a machine learning-based SER system is a large amount of high- quality data. You would need to assess whether there is sufficient data available for the specific use case you are targeting.

Computing resources: Machine learning algorithms can be computationally intensive, especially when dealing with large amounts of data. You would need to ensure that you have access to sufficient computing resources, such as powerful processors and GPUs, to train and run the SER system.

Integration with existing systems: You would need to assess whether the SER system can be easily integrated with existing systems in the target environment. For example, if the system is being deployed in a call center, it would need to be compatible with the call center software and hardware.

User training: If the SER system is being used by human operators, you would need to assess whether they can be trained to use the system effectively. This might involve providing training on how to interpret the output of the system and take appropriate action based on the detected emotions.

Legal and ethical considerations: Finally, you would need to consider any legal and ethical considerations associated with using an SER system. For example, you would need to ensure that the system is not biased towards any particular group or individual and that user privacy is protected.

4.6 Project Planning and Scheduling

Defining the project scope : The first step in any project planning process is to define the scope of the project. This would involve identifying the problem statement and the objectives of the project. For Speech Emotion Recognition, the objective would be to develop a machine learning model that can accurately recognize different emotions from speech data.

Identifying the stakeholders : The next step would be to identify the stakeholders involved in the project. This would include the project team members, the project sponsor, and any other stakeholders who may have an interest in the project outcome.

Developing a project plan : The project plan would outline the specific tasks and activities required to complete the project. This would include defining the project timeline, identifying the resources required, and determining the budget for the project.

Gathering data: The next step would be to gather the speech data required to train the machine learning model. This would involve collecting speech samples from various sources, including public datasets and user-generated content.

Preprocessing the data: Once the data has been collected, it would need to be preprocessed to remove any noise and ensure that it is in a format that can be used to train the machine learning model.

Training the machine learning model: The machine learning model would need to be trained using the preprocessed data. This would involve selecting the appropriate algorithm, determining the hyperparameters, and evaluating the model's performance.

Testing and validation: Once the machine learning model has been trained, it would need to be tested and validated to ensure that it can accurately recognize different emotions from speech data.

Deployment and maintenance: The final step would be to deploy the machine learning model and ensure that it is being used effectively. This would involve monitoring the model's performance and making updates as necessary to ensure that it continues to produce accurate results.

4.7 Software system Requirements

SYSTEM REQUIREMENS

HARDWARE REQUIREMENTS:

- System: Pentium i3 Processor.
- Hard Disk: 500 GB.
- Monitor: 15'' LED
- Input Devices: Keyboard, Mouse
- Ram: 4 GB

SOFTWARE REQUIREMENTS:

- Operating system: Windows 10.
- Coding Language: Python 3.8
- Web Framework: Flask

Chapter 5

System Design

5.1 Users of the System

Researchers: Researchers in the field of speech emotion recognition can use the system to test and evaluate their models and algorithms.

Developers: Developers can use the system to create applications that require emotion recognition from speech, such as virtual assistants or emotion-based music players.

Healthcare Professionals: Healthcare professionals can use the system to monitor the emotional state of patients, such as those with depression, anxiety, or post-traumatic stress disorder.

Educators: Educators can use the system to assess the emotional state of students during online learning or in-class lectures and provide tailored feedback accordingly.

5.2 Modularity criteria

Feature Extraction: The first module in an SER system is responsible for extracting features from speech signals that can be used for classification. This module should be designed to extract relevant features that are discriminative for different emotions. Commonly used features include Mel-frequency cepstral coefficients (MFCCs), pitch, energy, and formants.

Feature Selection: Once the features are extracted, the next module should select the most relevant features for classification. This module should be designed to reduce the dimensionality of the feature space and to eliminate irrelevant and redundant features.

Classification Algorithm: The third module should use a classification algorithm to classify the speech signal into different emotions. Various classification algorithms such as Support Vector Machines (SVM), k- Nearest Neighbors (k-NN), and Neural Networks (NN) can be used. The selection of the classification algorithm should be based on the performance of the algorithm on the given dataset.

Training and Testing: The fourth module should be designed to train thesystem using a labeled dataset and test the system using an unseen dataset. This module should be designed to avoid overfitting and to optimize the hyperparameters of the classification algorithm.

Integration: The final module should integrate all the modules together to form a complete system. This module should be designed to ensure that the system is scalable and can be easily modified to incorporate newfeatures or classification algorithms.

5.3 Design Methodologies

Data collection: The first step in any machine learning project is to gather a dataset of relevant information. In this case, we need a dataset of speech samples that includes a range of emotions, such as happiness, sadness, anger, fear, and neutral. This dataset can be collected from public databases or recorded specifically for this project.

Pre-processing: Once the dataset is collected, the speech signals need to be pre- processed to remove noise and irrelevant information. This can involve filtering, normalization, and resampling the speech signals to ensure they are all of a consistent quality.

Feature extraction: The next step is to extract relevant features from the pre- processed speech signals. This can include spectral features such as Mel-frequency cepstral coefficients (MFCCs), prosodic features such as pitch and energy, and temporal features such as zero-crossing rate and duration.

Feature selection: With a large number of features extracted, it is important to select the most relevant features for the classification task. This can be done using statistical methods or feature ranking techniques such as principal component analysis (PCA).

Classification: With the selected features, a machine learning algorithm can be trained to classify the emotional content of the speech signals. Common algorithms used for this task include support vector machines (SVM), random forests, and neural networks.

Evaluation: The final step is to evaluate the performance of the speech emotion recognition system. This can be done using metrics such as accuracy, precision, recall, and F1 score. The system can also be tested on new, unseen data to measure its generalization performance.

Fine-tuning: If the performance of the system is not satisfactory, the methodology can be refined by adjusting the feature extraction or selection techniques, trying different machine learning algorithms, or adding more data to the dataset.

5.5 Architecture diagrams

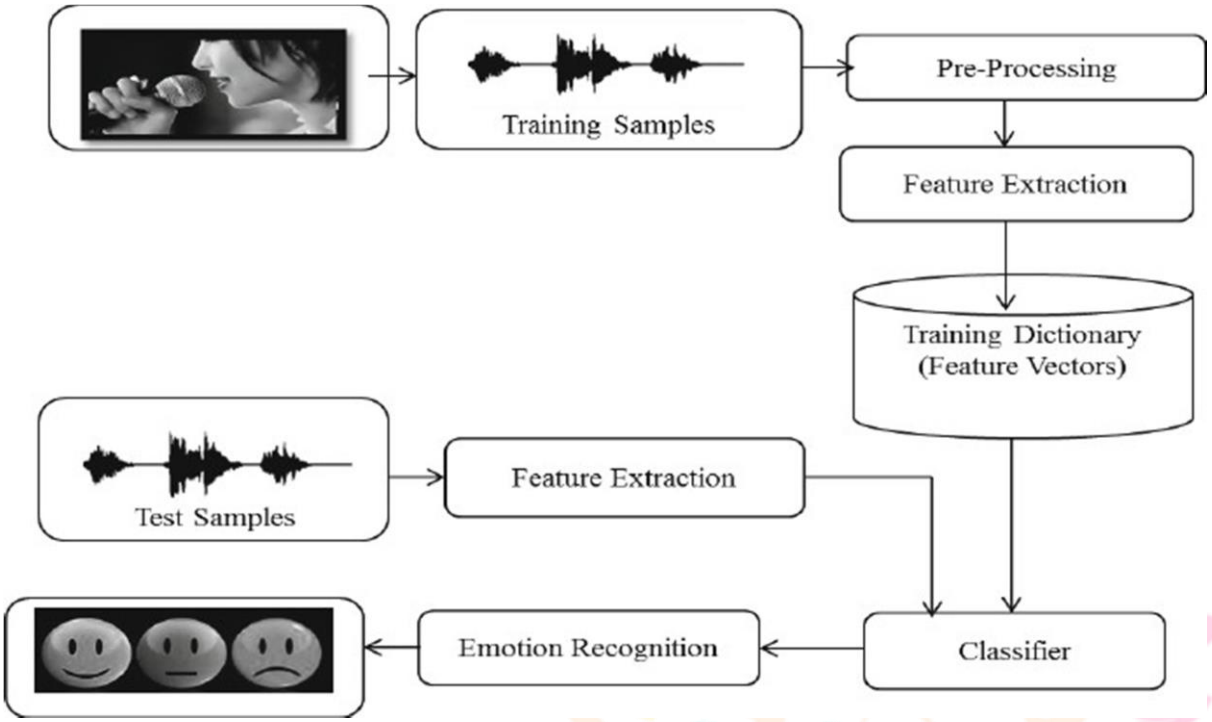


Figure 1: Architecture diagrams

5.5.1 Architecture User Interface Layouts Single Input Screen:

This layout involves a single input screen where users can upload or record a speech signal and receive feedback on the emotional content of the signal. The screen can include a waveform display of the speech signal, buttons for recording and playback, and a visual representation of the emotion detected.

Multi-Input Screen:

This layout involves a multi-input screen where users can upload or record multiple speech signals and receive feedback on the emotional content of each signal. The screen can include a list of the uploaded signals, a waveform display of each signal, and a visual representation ofthe emotion detected for each signal.

Real-time Analysis Screen:

This layout involves a real-time analysis screen where users can speak into a microphone and receive immediate feedback on the emotional content of their speech. The screen can include a waveform display of the speech signal, a visual representation of the emotion detected, and a text display of the detected emotion.

Comparison Screen:

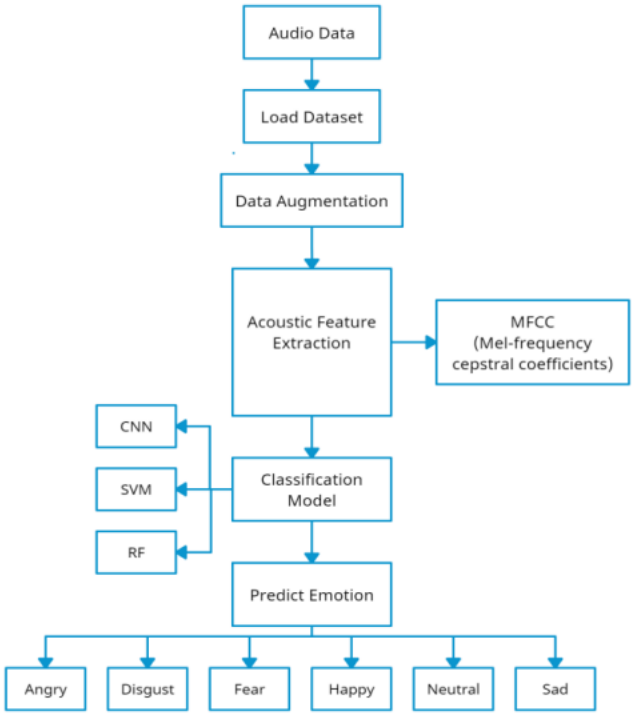
This layout involves a comparison screen where users can compare the emotional content of multiple speech signals. The screen can include a list of the uploaded signals, a waveform display of each signal, and a visual representation of the emotion detected for each signal, with the ability to compare the emotional content side-by-side.

Interactive Screen:

This layout involves an interactive screen where users can manipulate the emotional content of a speech signal and receive feedback on the changes. The screen can include sliders or other controls for adjusting the emotional content, a waveform display of the speech signal, and a visual representation of the emotion detected for each setting.

DataFlow Diagram:

In the first step the Data Set is loaded, from which one audio file is taken .Then that audio file is passed through Data Augmentation, where one without Augmentation, Noise ,High Speed, Low Speed, Stretch, Pitch are created .Then the Feature extraction is done from each data augmented audio file through MFCC.



Black Diagram:

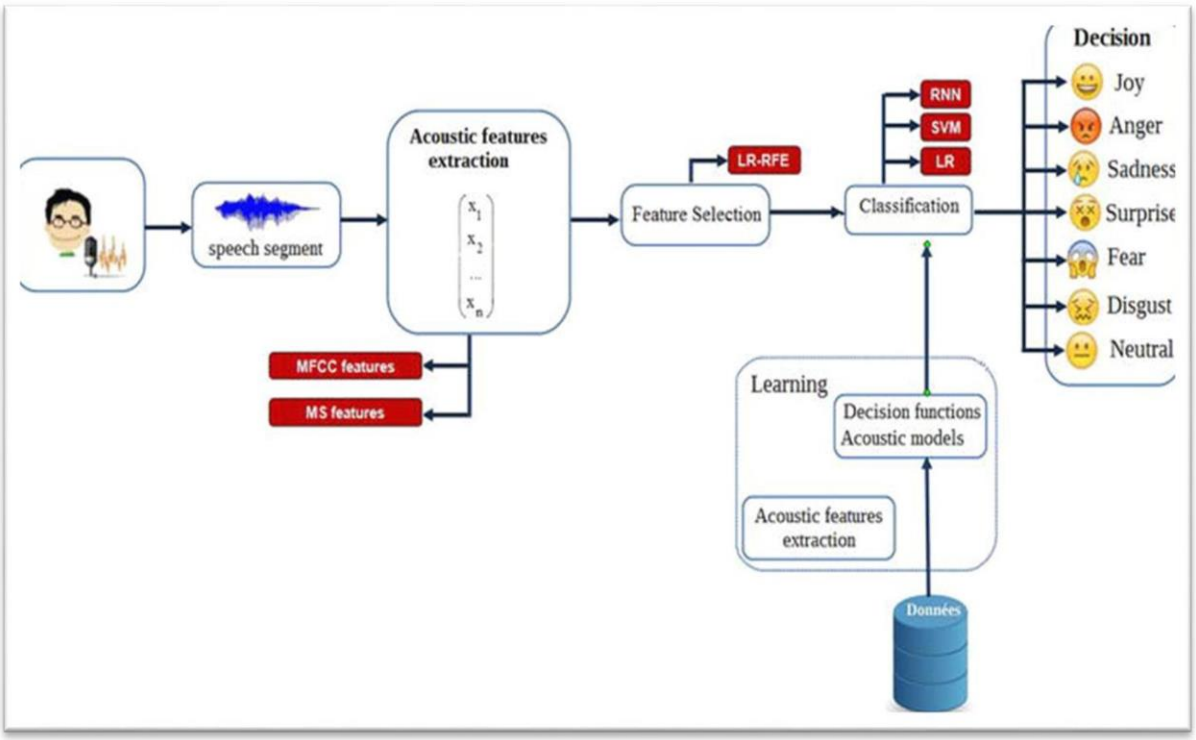


Figure 1: Block Diagram

CHAPTER 10

APPLICATION AND FUTURE ENHANCEMENT

Combining speech analysis with other biometric data such as facial expressions, gestures, and physiological signals can provide a more comprehensive understanding of emotional states.

Developing emotion recognition systems that can recognize emotions across different cultures and languages can improve the accuracy and applicability of these systems.

Developing real-time emotion recognition systems that can analyze emotions as they are expressed can be useful in situations where immediate feedback or intervention is necessary

Integrating speech emotion recognition with other modalities such as facial expressions, body language, and physiological signals can improve the accuracy of emotion detection.

Incorporating contextual information such as the speaker's age, gender, cultural background, and situational factors can improve the performance of emotion recognition models.

Transfer learning techniques can be used to train speech emotion recognition models on large-scale datasets and fine-tune them for specific applications.

Developing explainable AI models that can provide insights into how the emotion recognition system works and why certain decisions are made can increase user trust and transparency.

CONCLUSION

The TESS dataset that we considered is fine-tuned. Since it has noiseless data, it was easy for us to classify and feature the data. The classifier with utmost accuracy, AUC, F1 score, kappa, MCC is Random Forest Classifier. The classifier with the least accuracy and all the other terms is the decision tree classifier. We also mentioned why the decision tree classifier has low precision and the random forest classifier has high accuracy. The other classifiers that we analyzed, i.e., extra trees classifier, light gradient boosting machine, multi perceptron classifier, gradient boosting classifier, have the mid values in accuracy and other values. In conclusion, the random forest classifier is the most accurate algorithm and can be used in real-life scenarios to detect a person's emotions through his speech.

Rererences

Deng, Y., & Liu, M. (2017). Speech emotion recognition using deep neural network and extreme learning machine. IEEE Signal Processing Letters, 24(6), 781-785.

Kim, J., & Lee, J. (2018). Speech emotion recognition using convolutional neural network and transfer learning. Applied Sciences, 8(5), 699.

Li, X., Li, M., Yin, X., & Li, Y. (2020). Speech emotion recognition based on deep convolutional neural network with transfer learning. IEEE Access, 8, 29658-29666.

Jindal, G., Kumar, A., & Kumar, R. (2019). A survey on speech emotion recognition using machine learning techniques. International Journal of Speech Technology, 22(3), 469-488.

Ortega, J. D., Ochoa, S. F., & Serrano, J. I. (2020). Speech emotion recognition based on deep learning techniques: A review. Electronics, 9(11), 1906.

Zhao, M., Zheng, X., & Zhang, Y. (2019). A comparative study of speech emotion recognition using deep neural networks. IEEE Access, 7, 159642-159653.

Khan, M. M. U., Alam, M. M., & Begum, S. (2019). Speech emotion recognition using machine learning: A review. Journal of Ambient Intelligence and Humanized Computing, 10(9), 3459-3477.