International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 08 Issue: 06 | June - 2024 SJIF Rating: 8.448 ISSN: 2582-3930

Decoding Emotions: Machine Learning Approach to Speech Emotion Recognition

Manya Barua, Moulik Agwaral
Department of Artificial Intelligence and Data Science
Bhagwan Parhuram Institute of Technology, Rohini, Sector-17, New Delhi-11089

manyabarua@gmail.com, agrawalmoulik7@gmail.com

Dr. Varsha Sharma
Department of Artificial Intelligence and Data Science
Bhagwan Parshuram Institute of Technology, Rohini, sector-17, New Delhi-110089
varshasharma@bpitindia.com

Abstract—Speech Emotion Recognition (SER) stands at the forefront of human-computer interaction, offering profound implications for fields such as healthcare, education, and entertainment. This project report delves into the application of Machine Learning (ML) techniques for SER, aiming to discern the emotional content from speech signals.

The report begins with an overview of the significance of SER in various domains, emphasizing the need for accurate and robust emotion detection systems. Following this,a detailed exploration of the methodologies employed in SER is presented, encompassing feature extraction techniques, classification algorithms, and model evaluation metrics.

In the implementation phase, diverse ML algorithms such as Support Vector Machine (SVM), Convolutional Neural Networks (CNN), and Recurrent Neural Networks (RNN) are employed to classify emotional states from speech data. Feature sets including prosodic features, spectral features, and deep learning-based representations are extracted from the speech signals to capture nuanced emotional cues.

I. Introduction

In recent years, the intersection of artificial intelligence and human emotion has emerges as a captivating field of research with profound implications across various domains. Among these, Speech Emotions recognition (SER) stands as a pivotal area, leveraging the power of machine learning to decipher the intricate nuances of human expression through vocal cues. This project report delves into the exploration and implementation of SER using cutting-edge machine learning techniques, aiming to decode the emotional content embedded within spoken language.

The ability to discern emotions from speech holds immense significance, spanning from enhancing human-computer interaction to revolutionizing mental health diagnostics and beyond. With the advent of advanced computational algorithms and the availability of vast datasets, researchers have begun unraveling the complex patterns underlying emotional speech, paving the way for more nuanced and accurate recognition systems.

This research paper encapsulates the journey of developing a robust SER framework, encompassing various stages such as data pre-processing, feature extraction, model selection and performance evaluation. Through meticulous experimentation and analysis, insights are gleaned into the efficiency of different machine learning algorithms, their ability to capture subtle emotional cues, and the challenges encountered in real-world application scenarios.

Furthermore, this report also sheds light on the broader implications of SER, including its potential applications in fields such as human-computer interaction,, sentiments analysis, healthcare, education, and beyond. By bridging the gap between artificial intelligence and human emotion, SER not only unlocks new avenues for technological innovation but also fosters a deeper understanding of human behaviour and communication. Key Features:

- *User-Friendly Interface:* The app features an intuitive user interface designed to facilitate easy navigation and

interaction. With vibrant visuals and clear instructions, users can effortlessly navigate through the app's functionalities.

- *Audio Prediction:* Utilizing advanced machine learning models, the app analyzes audio inputs in real-time, predicting the underlying emotional states with impressive accuracy. Users can simply upload audio files in the '.wav' format, and the app swiftly processes the data to deliver insightful predictions.

- *Prediction History:* To enhance user experience and facilitate reflection, the app maintains a comprehensive history of predictions. Users can review past predictions, track their emotional journey, and gain deeper insights into their speech patterns and emotional expressions over time.

- *About The project frontend:* Curious users can delve into the app's background and development journey through the "About The App" section. Here, they can discover the inspiration behind the project, acknowledgments to contributors, and the technological framework powering the app's functionalities.

II. Literature Survey

A. Background

Due to its many uses in areas including affective computing, mental health evaluation, human-computer interaction, and personalized user experiences, speech emotion recognition, or SER, has attracted a lot of attention recently. Human communication is greatly impacted by emotions, which also have an impact on social relationships, decision-making, and general well-being. Handcrafted features and rule-based systems were the mainstays of traditional SER techniques, which frequently failed to adequately represent the complex and dynamic nature of human emotions.

But SER has advanced significantly with the introduction of deep learning and machine learning methods, allowing for more reliable and precise emotion identification from speech signals. Emotion identification is a hard endeavour, and machine learning models, especially neural networks, are good at extracting complex patterns and representations from data. These models are able to identify subtle emotional cues contained in vocal properties like pitch, intensity, rhythm, and spectral characteristics since they have been trained on vast datasets of labelled speech samples.

Notwithstanding the advancements in the subject, problems still exist in real-time and context-aware emotion recognition and in guaranteeing model generalization over a range of linguistic and cultural contexts and populations. Researchers from the fields of signal processing, machine learning, psychology, linguistics, and human-computer interaction must work together across disciplinary boundaries to address these issues.

B. Literature Survey

A comprehensive literature survey reveals a wealth of research contributions and methodologies employed in the domain of Speech Emotion Recognition using machine learning concepts. Several seminal works have laid the foundation for modern SER approaches, including:

1. *Data Sources*:

- Public Datasets: Utilize publicly available datasets specifically curated for speech emotion recognition research. Separate datasets are made for male and female voices for machine to learn how the pitch varies and also to learn the variation the emotions accordingly. Datasets are made manually to train the model according to the Indian accent so that the model can catch the emotion and identity it properly.

Online Platforms: Explore platforms like Kaggle, UCI Machine Learning Repository, and Zenodo for datasets related to speech emotion recognition tasks.

Data Collection: Collect and annotate your own dataset using crowd-sourcing platforms or in-house recordings to address specific research questions or domain requirements.

2. Data Annotation:

- Emotion Labelling: Annotate speech samples with categorical emotion labels such as happy, sad, angry, surprised, neutral, fear, and disgust based on perceptual judgments or self-reporting.

- Dimensional Annotations: Optionally, annotate emotional content along dimensional scales (e.g., valence, arousal) for nuanced analysis.

- Consistency Checks: Ensure inter-annotator agreement through reliability checks and consensus among annotators to maintain data quality.

3. *Preprocessing:*

- Audio Conversion: Convert audio files to a standardized format (e.g., WAV) with consistent sampling rates and bit depths to facilitate compatibility across tools and models.
- Normalization: Normalize audio signals to mitigate variations in volume and intensity, ensuring uniformity across the dataset.
- Feature Extraction: Extract relevant acoustic features from speech signals, including but not limited to the pitch of the voice, its tone, and energy. Its features are also depending on the manually made dataset that we used to train the model.

4. *Data Splitting*:

- Training, Validation, and Testing Sets: For training, validating, and evaluating the model, divide the dataset into mutually exclusive subsets for male and female manually made datasets which includes 100 to 200 voice notes of every possible emotion.
- Cross-Validation: Optionally, perform k-fold cross-validation to assess model performance across multiple splits and mitigate over-fitting.

5. *Data Augmentation* :

- Pitch Shifting: Perturb pitch to simulate variations in speaker characteristics and emotional expressiveness.
- Speed Perturbation: Modify speech rate to account for natural speech variability and environmental conditions.
- Noise Injection: Introduce background noise or environmental sounds to enhance model robustness against real-world acoustic conditions.

6. *Ethical Considerations*:

- Informed Consent Make sure that all data gathering practices follow ethical norms, which include getting participants' informed consent and protecting participants' privacy and confidentiality.
- Bias Mitigation: Using strategic sampling and data augmentation approaches, address potential biases in the dataset, including age, gender, and cultural biases.
- Data Integrity: Maintain data integrity and transparency by documenting data collection procedures, annotation protocols, and any pre-processing steps applied to the dataset.

III. Methodology and Experimentation

This aims to provide information on algorithms used in machine learning for speech emotion recognition. This document will give a comprehensive overview of our project, including the product point of view, user requirements, fundamental limitations, and an outline of the requirements. Furthermore, it will offer the particular specifications and features required for this project, including interface, functional requirements, and performance needs.

The scope of this document covers the project's whole life cycle. The final version of the software requirements, as decided upon by the clients and designers, is specified in this document. Ultimately, all capabilities may be traced from the SRS (Software Required Specification) to the product after the project's execution. Throughout the project's whole life cycle, the functionality, performances, constraints, interface, and dependability are all described in the document.

A machine learning (ML) model is used to create the voice emotion detection system. The implementation processes are similar to those of any other machine learning project, with the inclusion of extra steps for fine-tuning the model's performance. The flowchart provides an illustrated summary of the procedure. The data that is provided to the model during its development will help it learn, and the data will serve as the basis for all decisions and outcomes that the evolved model generates. A series of machine learning activities are performed on the gathered data in the second step, known as feature engineering. Numerous problems with data representation and quality are addressed by these approaches.

An algorithmic-based model is constructed in the third step, which is frequently regarded as the center of an ML project. This model trains itself to react to any new data it encounters by using an ML algorithm to learn about the data. Evaluating the built model's functionality is the last stage. The process of creating a model and assessing it is frequently repeated by developers in order to compare the effectiveness of various algorithms. Comparison findings aid in selecting the most suitable machine learning algorithm for the given task. .

In the current study, we proposed a speech emotion recognition (SER) system that classifies emotions using machine learning methods. The effectiveness of the emotion detection system can

have a significant impact on the application's overall performance in a variety of ways and offer a number of advantages over these applications' features. In order to better effectively identify speech perceptions based on emotions, this research provides a speech emotion detection system that improves over an existing system in terms of data, feature selection, and technique. .

CLASSIFIERS:

Classification is the process of recognizing, understanding, and organizing concepts and objects into predefined groups, sometimes known as sub-populations. These pre-categorized training datasets enable machine learning programs to classify future information into suitable and relevant categories using a variety of algorithms. Machine learning classification methods employ input training data to estimate the probability or likelihood that the subsequent data will fit into one of the predefined categories. Today's leading email service providers use one of the most popular uses of classification: separating emails into –spam and –non-spam categories. To put it briefly, pattern recognition is a type of classification.

Here, classification algorithms trained on the training data identify patterns in subsequent data sets, such as similar word or sentiment sequences, similar number sequences, and the like. We will look into the complicated workings of classification algorithms and learn how a text analysis program can carry out tasks such as sentiment analysis, which is employed to classify text that is unorganized according to the polarity of opinions (positive, negative, neutral, and so forth) .

SVM techniques create 3-dimensional classification models that go beyond X/Y predictive axes by classifying data and training models within extremely restricted degrees of polarity.

RANDOM FOREST CLASSIFIER:-

A number of decision trees develop into the classification algorithm known as the "Random Forest Classifier." In order to create uncorrelated forests, the algorithm employs allocation in the building process of each individual tree. The forest then utilizes its predictive capabilities to make precise decisions. The broad category of ensemble-based learning techniques includes random forest classifiers. They have shown to be incredibly successful in a range of disciplines, are easy to install, and operate quickly.

The building of several simple decision trees in the training stage and the majority vote (manner) among them in the classification stage constitute the fundamental idea of the random forest approach. This voting technique corrects for decision trees' unwanted potential to overfit training data, among other benefits. Random forests apply the general bagging strategy to each individual tree in the ensemble during the training phase. Using replacement, bagging repeatedly chooses a random sample from the training set and fits trees to these samples. No tree is ever maintained as it grows. A free parameter that can be easily learned automatically employing the so-called out-of-bag error is the number of trees in the ensemble.

Comparable to naıve Bayes and k-nearest neighbor algorithms, random forests are well-liked in part because of their ease of use and generally strong performance. But in contrast to the first two methods, random forests show some degree of unpredictable final trained model structure. This is an inevitable byproduct of tree building's random structure. As we will discuss in more detail soon, one of the main reasons that this feature of random forests can be problematic for regulatory reasons is that clinical adoption frequently necessitates a high degree of repeatability in terms of both the final algorithmic performance and the mechanics underlying a particular decision.

K-Neighbours-Classifier

It is difficult to express the idea of the k-nearest neighbor classifier more simply. This is an old proverb that is said in many different languages and civilizations. This indicates that we use the k-nearest neighbor classifier in our daily lives and in our decision-making.

Imagine that you run into a bunch of stylish, youthful, and active individuals. They discuss their friend Tony, who isn't present. What does Tony look like in your mind then? Yes, you do envision him as being fashionable, youthful, and athletic. If you find out that Tony resides in a neighborhood where conservative candidates are supported and where the annual income is over $200,000, what would you think? Do either of his neighbors earn more than $300,000 annually? How do you feel about Tony? You probably don't think of him as the underdog, and you might even think of him as a conservative.

The nearest neighbor classification method works by determining the specified number, or "k," of training samples that are closest to a fresh sample that needs to be classified. The new sample's label will be determined by analyzing its neighbors.

The number of neighbors that must be found for k-nearest neighbor classifiers is fixed and defined by the user. Additionally, there are neighbour learning algorithms that are

based on a set radius and have a variable number of neighbours based on the local density of points within the radius. In general, any metric measure can be used to measure the distance; the

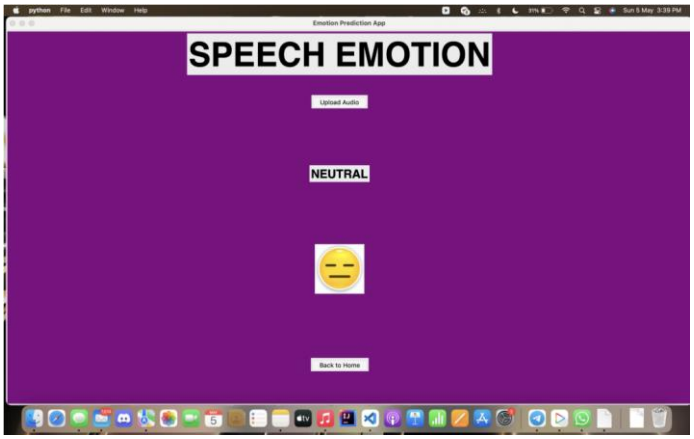

Fig. 1: output produced after giving the input

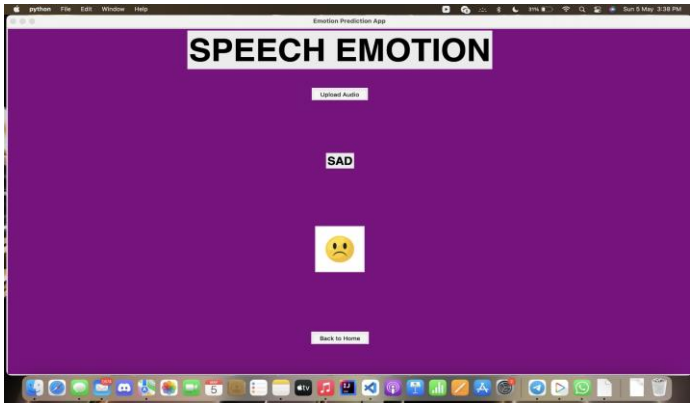

Fig. 2: output produced after giving sad input

Proach can be expanded inThe futureto handle multilingualism and emotion recognition. Additionally, it can be expanded conventional Euclidean distance is the most popular option. Since neighbors-based techniques only ”remember” all of the training data, they are referred to as non-generalizing machine learning techniques. A majority vote among the closest neighbors of the unidentified sample can be used to determine the classification.

Despite being one of the most straightforward machine learning algorithms, the k-NN algorithm has shown remarkable performance in a wide range of classification and regression tasks, such as image analysis and character recognition.

Among all machine learning algorithms, the k-nearest neighbor classification algorithm is one of the simplest. One kind of instance-based learning, sometimes known as lazy

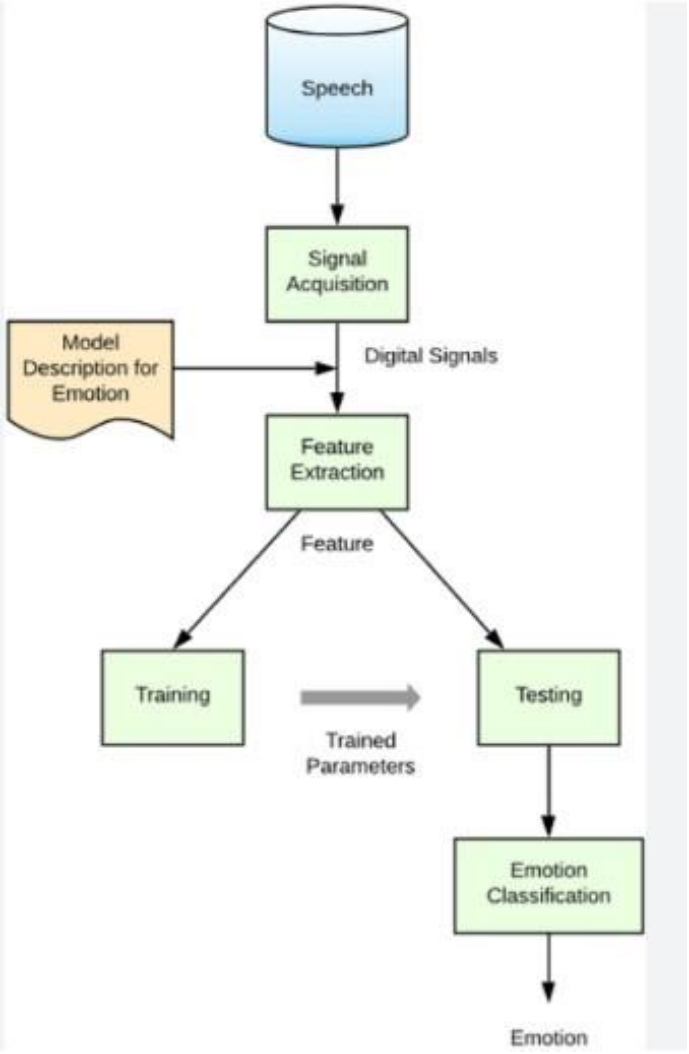

Fig. 3: block diagram.png

learning, is k-NN. Lazy learning, as used in machine learning, is a learning strategy that postpones generalizing the training set until a user queries the system. Conversely, eager learning occurs when the system typically makes generalizations from the training set prior to receiving queries. Put another way: When the real classification is being done, all computations are made and the function is only locally approximated.

IV. Conclusion and Future Work

In computer science, SER is a hot and interesting research area. The suggested system offers a cutting-edge SER algorithm with accurate real-time recognition. The suggested apto identify emotions at a minuscule level in relation to the situation.

Acknowledgment

We would like to express our sincere gratitude to everyone who helped us finish this research work,"Speech Emotion

Recognition using Machine Learning." First and foremost, we would like to sincerely thank our distinguished academic members for their essential advice and assistance throughout this project.

We would especially want to thank our department head, Dr. Varsha Sharma, for her knowledgeable counsel and support, which greatly improved the caliber of our work. Their observations have greatly influenced how we understand the topic and have helped us to improve our study techniques.

We also owe a debt of gratitude to our friends and peers for their support and encouraging words, which got us through many of the difficulties this project presented. Their resolute assistance acted as a continuous source of inspiration, encouraging us to pursue greatness.

Finally, we would want to thank our families from the bottom of our hearts for their unwavering support, love, and understanding. We were given the courage and tenacity to fully pursue our academic goals because of their steadfast support.

References

1. https://www.kaggle.com/code/jocelyndumlao/tess-emotion-recognition-deep-learning/
2. https://jpinfotech.org/speech-emotion-recognition-using-machine-learning/
3. https://blog.dataiku.com/speech-emotion-recognition-deep-learning/
4. https://www.irjet.net/archives/V10/11/IRJET-V10I1166.pdf
5. https://data-flair.training/blogs/python-mini-project-speech-emotion-recognition/
6. https://ijrpr.com/uploads/V3ISSUE5/IJRPR4210.pdf
7. https://www.mdpi.com/2076-3417/13/8/4750/
8. https://www.researchgate.net/publication/335360469_Speech_Emotion_Recognition_Using_Deep_Learning_Techniques_A_Review/
9. https://www.slideshare.net/SaniyaShaikh72/speech-emotion-recognition/
10. https://www.intechopen.com/chapters/65993/
11. https://medium.com/analytics-vidhya/speech-emotion-recognition-using-machine-learning-df31f6fa8404/
12. https://en.wikipedia.org/wiki/Speech_recognition/
13. https://simple.wikipedia.org/wiki/Speech_recognition/