**Speech Emotion Detection Using Hybrid Machine Learning**

**Abstract**

The current research aims to detect speech emotions by analyzing audio data. It uses a database from Kaggle with 1440 the wave speech files from different people. The MFCC method was applied for audio feature extraction, resulting in eight data sets. Outliers were identified and addressed using principal component analysis (PCA), and the MinMaxScaler method was used to normalize the data. To detect speech emotions, 5 machine learning algorithms have been chosen based on the review of the previous research papers. The selected machine learning models for detecting speech emotions have performed well, with Random Forest and Decision Tree models showing strong results. By suing these two models, two hybrid models, Stacking and Voting, were developed to combine the strengths of these models. The Stacking Hybrid Model achieved an impressive accuracy rate of 99.17% at 48000 Hz, surpassing previous research. The analysis highlights the importance of hybrid models in improving accuracy in speech emotion detection, promising future advancements in this area.

**Acknowledgement**

**Project Declaration**

**List of Abbreviations**

AI: Artificial Intelligence

IT: Information Technology

ML: Machine Learning

MFCC: Mel-Frequency Cepstrum Coefficient

STFT: Short-Time Fourier Transform

ANN: Artificial Neural Network

HCI: Human Computer Interaction

ASR: Automatic Speech Recognition

AUROC: Area under the Receiver Operating Characteristic Curve

LSTM: Long-Short Term Memory

SNR: Signal to Noise Ratio

CNN: Convolutional Neural Network

RNN: Recurrent Neural Network

**List of Figures**

[Figure 1 Gantt Chart for Research Planning 17](#_Toc183754186)

[Figure 2 Proposed Methodology 34](#_Toc183754187)

[Figure 3 Directories Under the RAVDESS Database 36](#_Toc183754188)

[Figure 4 Operation of MFCC 40](#_Toc183754189)

[Figure 5 Hybrid Model Design 45](#_Toc183754190)

[Figure 6 Environment Preparation for Artefact 46](#_Toc183754191)

[Figure 7 Data File Arrangement 47](#_Toc183754192)

[Figure 8 Audio Dataframe 48](#_Toc183754193)

[Figure 9 Data Visualization 48](#_Toc183754194)

[Figure 10 Speech Emotions and Distribution 49](#_Toc183754195)

[Figure 11 Data With Speech Emotions 49](#_Toc183754196)

[Figure 12 Feature-Extracted Data 50](#_Toc183754197)

[Figure 13 Feature Data 50](#_Toc183754198)

[Figure 14 Outlier Detection Result 51](#_Toc183754199)

[Figure 15 Normalized Data 51](#_Toc183754200)

[Figure 16 Model Queue 53](#_Toc183754201)

[Figure 17 Hybrid Model Design 53](#_Toc183754202)

[Figure 18 Performance of Models 55](#_Toc183754203)

[Figure 19 Result of the Optimum model 56](#_Toc183754204)

[Figure 20 Accuracy Correlation with Number of Features 59](#_Toc183754205)

**List of Tables**

[Table 1 Extractions from Previous Research 30](#_Toc183754206)

[Table 2 File Attributes and Nomenclature Process 37](#_Toc183754207)

[Table 3 Chosen Algorithms and Justification 44](#_Toc183754208)

[Table 4 Data Partition and Instances in Classes 52](#_Toc183754209)

[Table 5 Final Result of Speech Emotion Detection 54](#_Toc183754210)

[Table 6 Research Comparison 56](#_Toc183754211)

Table of Contents

[Chapter-1 Introduction 11](#_Toc183754100)

[1.1 Project Description 11](#_Toc183754101)

[1.1.1 Introduction 11](#_Toc183754102)

[1.1.2 Overview of Speech Emotion Detection 11](#_Toc183754103)

[1.1.3 Challenges in Speech Emotion Detection 11](#_Toc183754104)

[1.1.4 Problem Statement 12](#_Toc183754105)

[1.1.5 Scopes of Research on Hybrid Machine Learning for SED 12](#_Toc183754106)

[1.2 Research Particulars 13](#_Toc183754107)

[1.2.1 Research Questions 13](#_Toc183754108)

[1.2.2 Hypothesis 13](#_Toc183754109)

[1.2.3 Research Aim 13](#_Toc183754110)

[1.2.4 Research Objectives 13](#_Toc183754111)

[Chapter-2 Planning and Management of Research 15](#_Toc183754112)

[2.1 Research Issues 15](#_Toc183754113)

[2.1.1 Ethical Issues 15](#_Toc183754114)

[2.1.2 Legal Issues 15](#_Toc183754115)

[2.1.3 Professional Issues 16](#_Toc183754116)

[2.1.4 Financial Issues 16](#_Toc183754117)

[2.1.5 Social Issues 16](#_Toc183754118)

[2.2 Planning 17](#_Toc183754119)

[2.2.1 Gantt Chart 17](#_Toc183754120)

[2.2.2 Explanation of Tasks 17](#_Toc183754121)

[2.3 Required Resources 18](#_Toc183754122)

[Chapter-3 Review of Previous Papers 20](#_Toc183754123)

[3.1 Feature Extraction from Speech Data 20](#_Toc183754124)

[3.2 Speech Classification 22](#_Toc183754125)

[3.2.1 Applying Machine Learning 22](#_Toc183754126)

[3.2.2 Applying Deep Learning 25](#_Toc183754127)

[3.3 Speech Emotion Recognition System 27](#_Toc183754128)

[3.3.1 Applying Machine Learning 27](#_Toc183754129)

[3.3.2 Applying Deep Learning 29](#_Toc183754130)

[3.4 Research Gap Analysis 30](#_Toc183754131)

[3.4.1 Gaps in Previous Research 30](#_Toc183754132)

[3.4.2 Present Research Directions 32](#_Toc183754133)

[Chapter-4 Methodology 34](#_Toc183754134)

[4.1 Proposed Methodology 34](#_Toc183754135)

[4.1.1 Workflow 34](#_Toc183754136)

[4.1.2 Explanation 34](#_Toc183754137)

[4.2 Data Selection 36](#_Toc183754138)

[4.2.1 Data Source 36](#_Toc183754139)

[4.2.2 Data Description 36](#_Toc183754140)

[4.2.3 Speech File Attributes and Nomenclature Process 37](#_Toc183754141)

[4.2.4 Justification of Choice 38](#_Toc183754142)

[4.3 Tool Selection 38](#_Toc183754143)

[4.3.1 Python 38](#_Toc183754144)

[4.3.2 Jupyter Notebook 39](#_Toc183754145)

[4.3.3 Anaconda 39](#_Toc183754146)

[4.3.4 Advantages in Speech Processing and Classification 40](#_Toc183754147)

[4.4 Selection of Algorithms for Research 40](#_Toc183754148)

[4.4.1 Feature Extraction 40](#_Toc183754149)

[4.4.2 Outlier Treatment 41](#_Toc183754150)

[4.4.3 Emotion Classification 42](#_Toc183754151)

[4.4.4 Hybrid Model Selection and Design 44](#_Toc183754152)

[4.5 Evaluation Method 45](#_Toc183754153)

[Chapter-5 Analysis and Result 46](#_Toc183754154)

[5.1 Artefact Setup 46](#_Toc183754155)

[5.2 Processing Speech Database 47](#_Toc183754156)

[5.2.1 Data File Arrangement 47](#_Toc183754157)

[5.2.2 Audio Dataframe Preparation 47](#_Toc183754158)

[5.2.3 Data Visualization 48](#_Toc183754159)

[5.2.4 Creating Target Feature 48](#_Toc183754160)

[5.3 Audio Preprocessing 49](#_Toc183754161)

[5.3.1 Feature Extraction 49](#_Toc183754162)

[5.3.2 Feature Data Preparation 50](#_Toc183754163)

[5.3.3 Outlier Treatment 51](#_Toc183754164)

[5.4 Data Preparation 52](#_Toc183754165)

[5.5 Model Preparation 52](#_Toc183754166)

[5.5.1 Model Queue 52](#_Toc183754167)

[5.5.2 Design of Hybrid Model 53](#_Toc183754168)

[5.6 Speech Emotion Detection 53](#_Toc183754169)

[Chapter-6 Discussion and Conclusion 55](#_Toc183754170)

[6.1 Research Performance and Evaluation 55](#_Toc183754171)

[6.1.1 Research Performance 55](#_Toc183754172)

[6.1.2 Optimum Model 55](#_Toc183754173)

[6.1.3 Research Comparison and Improvement 56](#_Toc183754174)

[6.2 Discussion 57](#_Toc183754175)

[6.2.1 Discussion on Research Questions 57](#_Toc183754176)

[6.2.2 Discussion on Filling the Research Gaps 57](#_Toc183754177)

[6.2.3 Discussion on Challenges Faced and Mitigation 58](#_Toc183754178)

[6.2.4 Discussion on Hypothesis 59](#_Toc183754179)

[6.3 Conclusion 60](#_Toc183754180)

[6.3.1 Research Finding 60](#_Toc183754181)

[6.3.2 Limitations 61](#_Toc183754182)

[6.3.3 Future Scopes 61](#_Toc183754183)

[References 62](#_Toc183754184)

[Appendix 68](#_Toc183754185)

# Introduction

## Project Description

### Introduction

Speech emotion detection (SED) is an interdisciplinary field that combines elements of speech processing, psychology, and machine learning to identify and interpret human emotions from vocal cues. The integration of hybrid machine learning techniques has recently garnered attention due to its potential to enhance the accuracy and efficiency of emotion detection systems. Speech emotion detection using hybrid machine learning presents a promising avenue for addressing the inherent challenges in this field. By combining the strengths of various machine learning techniques, robust models can be developed which is capable of accurately detecting emotions from diverse speech data.

### Overview of Speech Emotion Detection

Speech emotion detection aims to recognize and classify human emotions based on speech signals. The fundamental components involve feature extraction from speech signals, followed by classification using machine learning algorithms. Traditional methods employed in SED include prosodic features (pitch, intensity, duration), spectral features (formants, mel-frequency cepstral coefficients), and linguistic features (textual content).

### Challenges in Speech Emotion Detection

Despite significant advancements, SED faces several challenges:

1. Variability in Speech Data: Emotional expressions can vary widely among individuals due to differences in accent, tone, and speaking style. This variability complicates the development of generalized models.

2. Ambiguity in Emotion Labels: Emotions are inherently subjective, and the labeling process can be inconsistent. Human annotators may disagree on the emotional content of a speech segment, leading to noisy training data.

3. Data Scarcity: High-quality, emotion-labeled speech datasets are limited. Collecting and annotating large-scale datasets is resource-intensive and time-consuming.

4. Real-time Processing: Implementing SED in real-time applications, such as virtual assistants or customer service bots, requires efficient algorithms that can process speech quickly without compromising accuracy.

### Problem Statement

The primary issue in SED is achieving high accuracy in emotion detection across diverse and real-world speech data. Traditional machine-learning approaches often struggle with the complexity and variability inherent in emotional speech. As a result, there is a pressing need to develop hybrid machine-learning models that can leverage the strengths of different algorithms to improve performance and robustness.

### Scopes of Research on Hybrid Machine Learning for SED

Hybrid machine learning models combine multiple algorithms to enhance the overall system's capabilities. In the context of SED, these models can integrate deep learning techniques with traditional classifiers, ensemble methods, or even incorporate external knowledge sources. The following avenues present promising scopes of research:

1. Combining Machine Learning Methods: The combination of Machine learning models excel at capturing complex patterns in speech data with the power of different algorithms invoked. Combining these models with traditional classifiers can improve feature representation and classification accuracy.
2. Ensemble Learning Techniques: Ensemble methods, such as Random Forests and Gradient Boosting Machines, can be employed to combine the predictions of multiple base models. This approach can mitigate the weaknesses of individual models and enhance overall performance. Research can focus on optimizing ensemble strategies for SED.
3. Transfer Learning and Domain Adaptation: Transfer learning can leverage pre-trained models on large datasets to improve emotion detection in smaller, domain-specific datasets. Domain adaptation techniques can address the issue of variability in speech data by adapting models to new speakers or recording conditions.
4. Multimodal Emotion Detection: Integrating speech with other modalities, such as facial expressions and physiological signals, can provide a more holistic understanding of emotional states. Hybrid models can be developed to fuse information from multiple sources, thereby improving detection accuracy.

## Research Particulars

### Research Questions

The research questions addressed in this study are as follows:

1. How can the hybrid model be designed to detect speech emotions with the highest performance?
2. How does this research improve upon existing models?

### Hypothesis

#### Null

**H0**: The detection accuracy of speech emotion does not depend upon the sample rate while extracting audio features.

#### Alternative

**H1**: The sample rate directly impacts the number of features extracted from the audio files.

**H2**: The detection accuracy of speech emotion depends upon the sample rate while extracting audio features.

### Research Aim

The aim of the research is to analyze the speech data for different speakers with different types if modulation and amplitudes and detect the inherited emotions using hybrid machine learning.

### Research Objectives

The objectives of the research are stated below in SMART format:

1. Literature Review

***Time***: 3 weeks

***Operation***: To review the existing research papers to gather the necessary knowledge on speech and audio processing, and speech emotion detection with the application of machine learning; This will help to choose algorithms to be applied in the present research.

***Outcome***: Gathering of ideas of existing research for methods and algorithms.

1. Data Collection

***Time***: 1 week

***Operation***: To select the database containing speech audio files with different types of emotions and study the data features such as file types, extensions, directories etc.

***Outcome***: Data will be obtained for research.

1. Data Preparation

***Time***: 2 weeks

***Operation***: To extract the audio features using the application of Mel spectrogram cepstral coefficient and remove outliers from the data to prepare if for emotion detection.

***Outcome***: Data will be prepared for speech emotion detection

1. Model Selection, Preparation and Application

***Time***: 4 weeks

***Operation***: To choose the machine learning models (from the literature review) which will be applied to the data for emotion detection; Tune the hyperparameters of the chosen models to prepare those compatible with machine learning; Apply those models to detect emotions from speeches.

***Outcome***: Speech Emotion Detection

1. Selection of Optimum Model by Performance Comparison

***Time***: 2 weeks

***Operation***: To compare the effectiveness of the models and select the optimum one to detect emotions with the highest accuracy and lower overfit; compare the performance of the optimum model with the existing approaches to determine the research improvement.

***Outcome***:

1. Preparing Final Documentation

***Time*** : 4 weeks

***Operation***: To proceed to documentation and complete the final project report with all implications made and outputs obtained with evaluation.

***Outcome***: Final Project Repossrt