**A**

**Major Project Report**

**On**

“SPEECHSENTIO: AI – POWERED SPEECH THERAPY WITH EMOTION ANALYSIS”

Submitted in partial fulfillment of the

Requirements for the award of the degree of

**Bachelor of Technology**

**In**

**Computer Science & Engineering –**

**Artificial Intelligence & Machine Learning**

**By**

**Behara Amulya – 20R21A6609**

**Jasthi Pavithra – 20R21A6621**

**Kunapulli Abhinay – 20R21A6627**

**Vasamsetty Lohitha – 20R21A6652**

Under the guidance of

**G. Aishwarya**

**Assistant Professor**

**Department of Computer Science & Engineering –**

**Artificial Intelligence & Machine Learning**

****

**2024**



**Department of Computer Science & Engineering-**

**Artificial Intelligence & Machine Learning**

**CERTIFICATE**

This is to certify that the project entitled **“SpeechSentio: AI – powered Speech Therapy with Emotion Analysis”** has been submitted by **Behara Amulya (20R21A6609), Jasthi Pavithra (20R21A6621), Kunapuli Abhinay (20R21A6627)** and **Vasamsetty Lohitha (20R21A6652)** in partial fulfillment of the requirements for the award ofdegree of Bachelor of Technology in Computer Science and Engineering – Artificial Intelligence & Machine Learning from Jawaharlal Nehru Technological University, Hyderabad. The results embodied in this project have not been submitted to any other University or Institution for the award of any degree or diploma.

**Internal Guide** **Head of the Department**

**External Examiner**

**i**



**Department of Computer Science & Engineering-**

**Artificial Intelligence & Machine Learning**

**DECLARATION**

We hereby declare that the project entitled **“Speech Sentio: AI – powered Speech Therapy with emotion analysis”** is the work done during the periodfrom **January 2024 to April 2024** and is submitted in partial fulfillment of the requirements for the award of degree of Bachelor of Technology in Computer Science and Engineering – Artificial Intelligence & Machine Learning from Jawaharlal Nehru Technology University, Hyderabad. The results embodied in this project have not been submitted to any other university or Institution for the award of any degree or diploma.

**Behara Amulya 20R21A6609**

**Jasthi Pavithra 20R21A6621**

**Kunapuli Abhinay 20R21A6627**

**Vasamsetty Lohitha 20R21A6652**

**ii**



**Department of Computer Science & Engineering-**

**Artificial Intelligence & Machine Learning**

**ACKNOWLEDGEMENT**

The satisfaction and euphoria that accompany the successful completion of any task would be incomplete without the mention of people who made it possible, whose constant guidance and encouragement crowned our efforts with success. It is a pleasant aspect that we now have the opportunity to express our guidance for all of them.

First of all, we would like to express our deep gratitude towards our internal guide

**G. AISHWARYA, Assistant Professor, Department of CSE(AI&ML)** for his support in the completion ofour dissertation. We wish to express our sincere thanks to **Dr.K.SAI PRASAD,** HOD, Dept. of CSE- AIML and principal **Dr. K. SRINIVAS RAO** for providing the facilities tocomplete the dissertation.

We would like to thank all our faculty and friends for their help and constructive criticism during the project period. Finally, we are very much indebted to our parents for their moral support and encouragement to achieve goals.

**Behara Amulya 20R21A6609**

**Jasthi Pavithra 20R21A6621**

**Kunapuli Abhinay 20R21A6627**

**Vasamsetty Lohitha 20R21A6652**

**iii**



**Department of Computer Science & Engineering**

**ABSTRACT**

The study delves into the intricate realm of stuttering detection, focusing on the complexities of this speech disorder and its integration with emotion recognition. It outlines the difficulty level of stuttering and considers contemporary techniques for detecting and measuring speech disfluencies, including signal processing methods and machine learning algorithms. Acknowledging the convoluted relationship between stammer patterns and emotional expressions, the research emphasizes the importance of including emotion identification in the stuttering detection framework. The study investigates different techniques for analyzing speech sounds to accurately identify patterns of disfluency, including examining visual representations of speech (spectrograms), analyzing the frequency content (cepstral analysis), and using powerful statistical methods (machine learning algorithms) to categorize these patterns. Integrating emotion recognition into stuttering detection involves identifying emotions from speech cues like prosody and intonation, and fusing features from both domains to develop robust models. Multimodal approaches, incorporating facial expressions and physiological signals, further enhance emotion recognition and its integration with stuttering detection. This investigation adopts a comprehensive perspective to illuminate the complexities of speech disorders. The aim is to leverage these insights to develop more efficacious intervention strategies applicable in clinical and therapeutic environments.

**iv**

**APPENDIX-1**

**LIST OF FIGURES**

**LIST OF FIGURES**

|  |  |  |
| --- | --- | --- |
| **Fig No** | **Description of Figure** | **Page No** |
| **F**i**g 1** | **McCulloch-Pitts Model** | **130** |
| **Fig 2** | **Rosenblatt’s Perceptron** | **130** |
| **Fig 3** | **Multi-layer Perceptron** | **131** |
| **Fig 4** | **Architecture of proposed system** | **135** |
| **Fig 5** | **Workflow of Stutter detection process** | **136** |
| **Fig 6** | **Workflow of Emotion Analysis process** | **138** |
| **Fig 7** | **Use Case Diagram** | **140** |
| **Fig 8** | **Class diagram for Emotion Analysis model** | **141** |
| **Fig 9** | **Class diagram for Stutter Detection model** | **142** |
| **Fig 10** | **Class diagram of a system that uses emotion data to identify a stutter type** | **142** |
| **Fig 11** | **Sequence diagram of Speechsentio** | **143** |
| **Fig 12** | **Activity Diagram of Speechsentio** | **144** |
| **Fig 13** | **SpeechSentio speech pattern feedback via user-uploaded audio** | **166** |
| **Fig 14** | **Audio Analysis Report for “WomenWhoStutter\_97\_82.wav”** | **167** |
| **Fig 15** | **Audio Analysis Report for “WomenWhoStutter\_99\_163.wav”** | **167** |
| **Fig 16** | **Audio Analysis Report for “03-01-07-01-01-01-04.wav” Fused Image** | **167** |
| **Fig 17** | **Confusion matrix for Decision Tree, KNN and MLP** | **168** |
| **Fig 18** | **Precision-Recall Curve for MLP** | **168** |
|  |  |  |

**APPENDIX-2**

**LIST OF TABLES**

|  |  |  |
| --- | --- | --- |
|  | **LIST OF TABLES** |  |
| **Table No** | **Description of Figure** | **Page No** |
| **Table 2.2** | **Comparison table** | **93** |
| **Table 2.3** | **Work Evaluation table** | **99** |
| **Table 4.1.1** | **Types of stutters and Descriptions** | **137** |
| **Table 6.1** | **Accuracies obtained on ML algorithms** | **169** |

**vi**

**APPENDIX-3**

**LIST OF ABBREVIATIONS**

**ABBREVIATIONS**

|  |  |
| --- | --- |
|  |  |
| **MFCC**  **RNN** | **Mel-Frequency Cepstral Coefficients**  **Recurrent Neural Network** |
| **CNN** | **Convolutional Neural Network** |
| **KNN** | **K-Nearest Neighbors** |
| **PET** | **Stochastic Gradient Descent** |

**vii**

# 

# APPENDIX-4

# REFERENCES

**REFERENCES**

1. Sadeen Alharbi, Madina Hasan, Anthony J H Simons, Shelagh Brumfitt, and Phil Green, 2017, ‘Stuttering from Transcripts: A Comparison of HELM and CRF Approaches’, eprints.whiterose.ac.uk, Vol. 1, pp. 1-11, 2017.
2. Noeth, E., Wittenberg, T., Decher, M., & Dietrich, S. (2000). Automatic stuttering recognition using hidden Markov models. In Proceedings of the International Conference on Spoken Language Processing (ICSLP) (pp. 753-756).
3. Kourkounakis, T., Hajavi, A., & Etemad, A. (2020). Detecting Multiple Speech Disfluencies Using a Deep Residual Network with Bidirectional Long Short-Term Memory. In 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 3347-3351). Institute of Electrical and Electronics Engineers (IEEE).
4. Al-Qatab, B. A., & Mustafa, M. B. (2021). Classification of Dysarthric Speech According to the Severity of Impairment: An Analysis of Acoustic Features. IEEE Access, 9, 18183-18194. Doi:10.1109/ACCESS.2021.3053335
5. Lalitha, S., Tripathi, S., & Gupta, D. (2019). Enhanced speech emotion detection using deep neural networks. International Journal ofSpeech Technology, 22(3), 497–510. Doi:10.1007/s10772-018-09572-8
6. H. Aouani and Y. B. Ayed, “Speech Emotion Recognition with deep learning,” Procedia Computer Science, vol. 176, pp. 251-260, 2020.
7. I, Husbaan. (2022). Speech Emotion Recognition System Using Machine Learning. International Journal of Research Publication and Reviews, 3(5), pp. 2869-2880.
8. Bhatia, G., Saha, B., Khamkar, M., Chandwani, A., & Khot, R. (2021). Stutter Diagnosis and Therapy System Based on Deep Learning. International Journal of Research Publication and Reviews, 5(1), 1-8.
9. Barda, S. (2019). Recognition of rate of stuttering in patients having speech disorders. International Journal of Research Publication and Reviews, 3(1), 1-6.
10. Mahendran,M.,Visalakshi, S., & Balaji, S. (2021). Dysarthria detection using CNN and MFCC feature extraction. International Journal of Research Publication and Reviews, 2(2), 1-7 .

**viii**

1. Mustaqeem and S. Kwon, “A CNN-Assisted enhanced audio signal processing for speech emotion recognition,” Sensors, vol. 20, no. 1, p.183, 2019.
2. M. Ghai, S. Lal, S. Duggal, and S. Manik, “Emotion recognition on speech signals using machine learning,” Mar. 2017, doi: 10.1109/icbdaci.2017.8070805.
3. Apeksha Aggarwal, Akshat Srivastava, Ajay Agarwal, Nidhi Chahal, Dilbag Singh, Abeer Ali Alnuaim, Aseel Alhadlaq and Heung-No Lee, “Two-Way feature extraction for speech emotion recognition using deep learning,” Sensors, vol. 22, no. 6, p. 2378, Mar. 2022,1-11.
4. T. Kourkounakis, A. Hajavi, and A. Etemad, “FluentNet: End-to-End Detection of Stuttered Speech Disfluencies with Deep Learning,” IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 29, pp. 2986–2999, Jan. 2021, 1-13.
5. Manas Jain, Shruthi Narayan, Pratibha Balaji, Bharath KP, Abhijit Bhowmick, Karthik R, Rajesh Kumar Muthu, “Speech Emotion Recognition using Support Vector Machine,” arXiv:2002.07590 [cs, eess], Feb. 2020, doi: <https://arxiv>.org/abs/2002.07590, 1-6.
6. Suwon Shon, Pablo Brusco, Jing Pan, Kyu J. Han, Shinji Watanabe “Leveraging Pre-trained Language Model for Speech Sentiment Analysis” INTERSPEECH 2021 30 August – 3 September
7. Bagus Tris Atmaja, Akira Sasou, “Sentiment Analysis and Emotion Recognition from Speech Using Universal Speech Representations”, 24 August 2022.
8. Shakeel, A. Sheikh, Md Sahidullah , Fabrice Hirsc, Slim Ouni, “StutterNet: Stuttering Detection Using Time Delay Neural Network” , Jan. 2021
9. Colin Lea, Zifang Huang, Jaya Narain, Lauren Tooley, Dianna Yee, Tien Dung Tran, “From User Perceptions to Technical Improvement: Enabling People Who Stutter to Better Use Speech Recognition”, April 23–28, 2023
10. Phani Bhushan S, Vani H Y, D K Shivkumar, “Stuttered Speech Recognition using Convolutional Neural Networks”, 2021
11. Badshah, A., Lee, S., & Kim, J. (2017). Speech Emotion Recognition from Spectrograms with Deep Convolutional Neural Network. International Journal of Research in Engineering and Technology, 6(6), 239-244.

**ix**

1. Li, S., Deng, L., & Huang, J. T. (2013). Hybrid Deep Neural Network – Hidden Markov Model Based Speech Emotion Recognition. International Journal of Research Publication and Reviews, 6(3), 312- 317.
2. Mahesha, P., & Vinod, D. S. (2016). Automatic Segmentation and Classification of Dysfluencies in Stuttering Speech. International Journal of Research Publication and Reviews, 3(4), 12-20.
3. Mahesha, P., & Vinod, D. S. (2017). LP-Hilbert Transform Based MFCC for Effective Discrimination of Stuttering Dysfluencies. International Journal of Research Publication and Reviews, 4(3), 2564-2571.
4. Narendra, N.P., & Alku, P. (2019). Dysarthric speech classification from coded telephone speech using glottal features. Speech Communication, 110, 47-55.
5. Zhao, J., Mao, X., & Chen, L. (2018). Speech emotion recognition using deep 1D & 2D CNN LSTM networks. Biomedical Signal Processing and Control, 47, 312-323.
6. Issa, D., Demirci, M. F., & Yazici, A. (2020). Speech emotion recognition with deep convolutional neural networks. Biomedical Signal Processing and Control, 59, 101894.
7. Alif Bin Abdul Qayyum, Asiful Arefeen, & Celia Shahnaz. (2019). Convolutional Neural Network (CNN) Based Speech-Emotion Recognition. International Journal of Research Publication and Reviews, 6(4), 122-130.
8. Afroz, F., & Koolagudi, S. G. (2019). Recognition and Classification of Pauses in Stuttered Speech using Acoustic Features. International Journal of Research Publication and Reviews, 6(2), 12-20.
9. Bhushan, P., Shivkumar, D. K., Vani, H. Y., & Sreeraksha, M. R. (Year). Stuttered Speech Recognition using Convolutional Neural Networks. International Journal of Research Publication and Reviews, Volume 9(Issue 12), pp. 250-254.

**x**

**INDEX**

**Certificate i**

**Declaration ii**

**Acknowledgement iii**

**Abstract iv**

**List of Figures v**

**List of Tables vi**

**List of Abbreviations vii**

**References viii**

**Chapter 1**

**Introduction**  1

* 1. Overview 1

1.2 Purpose of the project 1

1.3 Motivation 1

**Chapter 2**

**Literature Survey 2**

2.1 Existing System 2

2.2 Comparison table 93

2.3 Work Evaluation table 99

**Chapter 3**

**Proposed System**  125

3.1 Proposed System 125

3.2 Advantages of Proposed System 125

3.3 System Requirements 126

3.3.1 Software Requirements 126

3.3.2 Hardware Requirements 126

3.3.3 Implementation Technologies 127

3.3.4 Introduction to MLP 129 137

**Chapter 4**

**System Design** 134

4.1Proposed System architecture 133

4.1.1 Stutter detection module 135

4.1.2 Emotion recognition module 137

4.2 UML diagrams 140

4.2.1 Use Case Diagram 140

4.2.2 Class Diagram 141

4.2.3 Sequence Diagram 143

4.2.4 Activity Diagram 143

**Chapter 5**

**Implementation** 145

5.1 Brief Introduction of Implementation 145

5.1.1 Stutter Classification 145

5.1.2 Speech Emotion Recognition 146

5.1.3 Deployment of Speechsentio 147

5.2 Source Code 148

index.html 148

style.css 150

result.html 156

app.py 159

**Chapter 6**

**Results** 166

**Chapter 7**

**Conclusion** 170

**Future Enhancements and Discussions** 170

# CHAPTER 1

# INTRODUCTION

* 1. **OVERVIEW**

This initiative tackles communication hurdles, specifically stuttering, by exploring the ever-changing landscape of speech therapy. The objective of it is to establish better methods for recognize and address these challenges. Given the complexities of stuttering, this project dives into groundbreaking methods in speech therapy fueled by recent technological progress. This project tackles stuttering by looking at recent discoveries in speech analysis and fresh research. The aim is to create more reliable, flexible, and tailored methods for spotting and sorting these speech difficulties. The project also investigates incorporating technology that can pick up on emotions as a whole new approach to stuttering treatment. This could significantly improve how we diagnose stuttering and develop treatment plans.

* 1. **Purpose of the project**

The purpose of this project is twofold: firstly, this project aims to enhance the quality of life for people who struggle with speech difficulties in addition to using cutting-edge technology to further the knowledge and treatment of stuttering. By carefully examining speech data and incorporating emotion recognition technology, the project seeks to improve the precision of diagnosis, customize therapeutic interventions to meet the needs of each patient, and establish a nurturing atmosphere that encourages motivation and speeds up speech therapy advancements. The initiative aims to open the door for more effective, accessible, and customized methods of stuttering detection and therapy by embracing technology innovations and interdisciplinary collaboration.

* 1. **Motivation**

This project’s impetus comes from the realization of the ongoing difficulties with traditional approaches to stuttering identification and treatment. Given that stuttering is a complicated and multidimensional speech barrier, it is imperative to investigate novel approaches that can be tailored to the specifics of each person’s experience. The project aims to transform speech therapy by utilizing advances in signal processing, emotion identification technologies, and computer algorithms. This will empower individuals to overcome obstacles to communication and reach their maximum potential.

**CHAPTER 2**

**LITERATURE SURVEY**

An extensive literature survey has been conducted by studying existing systems of Speech Emotion Recognition and Stutter Detection. A good number of research papers, journals, and publications have also been referred before formulating this survey.

* 1. **Existing Systems**

|  |
| --- |
| **1** |
| **Reference in APA format** | |  | | | | |
| **URL of the Reference** | | **Authors Names and Emails** | | | **Keywords in this Reference** | |
| <https://ieeexplore>.ieee.org/document/9053893 | | [Tedd Kourkounakis](https://ieeexplore.ieee.org/author/37088481517)  [Amirhossein Hajavi](https://ieeexplore.ieee.org/author/37088482795)  [Ali Etemad](https://ieeexplore.ieee.org/author/37087028037) | | | Speech, stuttering, disfluency, deep learning, residual network, LSTM | |
| **The Name of the Current Solution (Technique/ Method/ Scheme/ Algorithm/ Model/ Tool/ Framework/ ... etc )** | | **The Goal (Objective) of this Solution & What is the problem that need to be solved** | | | **What are the components of it?** | |
| Deep residual network with bidirectional long short-term memory layers. | | The goal of this solution is to detect and classify different forms of stutters. The problem that needs to be solved is that there is minimal data and research on the identification and classification of stuttered speech. | | | Bidirectional long short- term memory (Bi-LSTM), spectrogram feature vectors, batch normalization and ReLu activation functions, root means square propagation (RMSProp) optimizer, softmax loss function, Nvidia 1080 Ti GPU | |
| **The Process (Mechanism) of this Work; Means How the Problem has Solved & Advantage & Disadvantage of Each Step in This Process** | | | | | | |
| The paper’s mechanism involves data collection, feature extraction, deep learning model training, and performance evaluation to detect and classify stutter disfluencies.   |  |  |  |  | | --- | --- | --- | --- | |  | **Process Steps** | **Advantage** | **Disadvantage (Limitation)** | | 1 | Speech samples are collected from the University College London’s Archive of Stuttered Speech (UCLASS) dataset. The dataset contains samples of monologues from participants with known stuttered speech impediments. Each recording is manually annotated for one of seven stutter disfluencies. | Using a real-world dataset with manual annotations provides a diverse and challenging dataset for training and evaluation. | The dataset may have limitations in terms of size and diversity, and manual annotation can be time-consuming. | | 2. | Spectrogram features are extracted from audio clips. Spectrograms represent audio signals in a format suitable for deep learning | Spectrograms are commonly used in speech analysis, and they capture the temporal and spectral characteristics of the speech signal. | Spectrograms may not capture all relevant information, and the choice of spectrogram parameters (e.g., window size) can impact results. | | 3. | A deep residual neural network is used for feature embedding. The network architecture includes convolutional layers with batch normalization and ReLU activation functions. | Deep residual networks are known for their ability to capture complex features. The architecture can learn stutter-specific features effectively. | Training deep networks can be computationally expensive and requires a large amount of data. | | 4. | The learned feature embeddings are processed by bidirectional LSTM layers. Each LSTM layer consists of 512 bidirectional LSTM units. | LSTM layers are effective in modeling sequential data. Bidirectional LSTMs capture context from both past and future embeddings. | The choice of hyperparameters (e.g., LSTM units, dropout rates) can affect model performance. | | 5. | The model is trained using TensorFlow’s Keras API with a root mean square propagation (RMSProp) optimizer and softmax loss function. Leave-one-subject-out (LOSO) cross-validation is used for rigorous testing. | Cross-validation ensures robust evaluation. The use of bidirectional LSTMs and deep networks contributes to the model’s performance. | Training deep learning models can be time-consuming and may require significant computational resources. | | 6. | The model's performance is evaluated in terms of accuracy and miss rate for each class of stutter disfluency. | The evaluation provides insights into the model’s ability to detect and classify different stutter types. | Evaluating the model’s performance can be challenging due to the need for diverse and well-annotated datasets. | | | | | | | |
| **Major Impact Factors in this Work** | | | | | | |
| |  |  |  |  | | --- | --- | --- | --- | | **Dependent Variable** | **Independent Variable** | **Moderating variable** | **Mediating (Intervening ) variable** | | **Stutter Type Classification Outcome:**  The outcome variable involves the categorization of distinct stutter types, such as sound repetition, word repetition, phrase repetition, revision, interjection, or prolongation, determined through the proposed deep residual network and bidirectional long short-term memory (LSTM) model. | **Acoustic Spectrogram Feature Vectors and Model Architecture:**  The independent variables encompass acoustic features, specifically spectrogram feature vectors derived from audio clips. Additionally, the architectural components of the deep learning model, including the deep residual network and bidirectional LSTM layers, contribute to the independent variables. | **Dataset Magnitude and Severity of Stutter Impairment:**  The moderating variables involve the scale of the dataset, characterized by the number of participants and audio samples, and the severity of stutter in participants. These factors may modulate the relationship between independent and dependent variables. | **Learned Feature Embeddings and Residual Network:**  The mediating variables consist of the feature embeddings acquired through the deep residual network. These embeddings serve as an intermediary step in capturing stutter-specific characteristics, addressing challenges such as the vanishing gradient problem during model training. | | | | | | | |
| |  | | --- | | **Relationship Among The Above 4 Variables in This article** | | In this study, the dependent variable is the classification outcome for different stutter types, influenced by acoustic spectrogram feature vectors and a deep residual network with bidirectional long short-term memory layers. The dataset, with various stutter severity levels, moderates the classification effectiveness. The mediating variable, represented by learned feature embeddings, acts as a crucial link between input features and classification. The interplay among these variables highlights the importance of dataset characteristics, model architecture, and quality of learned features in achieving robust stutter type classification based solely on acoustic features. | | | | | | | |
| **Input and Output** | | | **Feature of This Solution** | | | **Contribution & The Value of This Work** |
| |  |  | | --- | --- | | **Input** | **Output** | | Audio speech signals | Classification of different types of stutter disfluencies | | | | This solution integrates deep learning, residual networks, and bidirectional LSTMs to classify various stutter disfluencies from audio speech signals. It achieves state-of-the-art performance and holds potential for future research, including multi-class classification. Additional feature selection methods may further enhance results. | | | The contribution of this work lies in the development of a robust system for detecting and classifying various types of stutter disfluencies using deep learning, residual networks, and bidirectional LSTMs. This methodology offers a significant advancement in the field of speech disfluency analysis by avoiding the reliance on language models or ASR, making it more efficient and accurate. The system’s potential applications in therapy, speech monitoring, and improving presentation skills hold the promise of positively impacting the lives of individuals with stuttering impediments. |
| **Positive Impact of this Solution in This Project Domain** | | | | **Negative Impact of this Solution in This Project Domain** | | |
| This solution holds the potential to greatly improve the diagnosis and therapy of stuttering, benefiting millions affected by this speech impediment. By offering robust and efficient stutter detection through acoustic features, it enhances early intervention and therapeutic success rates. Its application in speech analysis can contribute to a higher quality of life for individuals with stuttering disorders. | | | | While this solution offers valuable advancements, its reliance on acoustic features alone may lead to occasional misclassification, particularly for more complex stutter types. Additionally, it may not address gender, accent, and speech rate variations, limiting its effectiveness in diverse populations. The potential for false negatives in classification could impact the accuracy of stutter detection. | | |
| **Analyse This Work By Critical Thinking** | | | **The Tools That Assessed this Work** | | | **What is the Structure of this Paper** |
| This work presents an innovative approach to stutter detection by focusing solely on acoustic features and avoiding ASR, enhancing efficiency. However, its performance limitations, such as occasional misclassification and sensitivity to variations, warrant further exploration for robust real-world applications. | | | TensorFlow, Keras, Nvidia CUDA Toolkit, Librosa, Scikit-learn, Jupyter Notebook, Audacity, MATLAB, PyTorch, GitHub | | | 1. Abstract 2. Introduction 3. Related Work 4. Proposed Methodology 5. Experiment Setup And Results 6. Conclusion and Future work 7. Acknowledgements |
| **Diagram/Flowchart** | | | | | | |
|  | | | | | | |

|  |
| --- |
| **2** |
| **Reference in APA format** | |  | | | | |
| **URL of the Reference** | | **Authors Names and Emails** | | | **Keywords in this Reference** | |
| <https://www>.sciencedirect.com/science/article/pii/S1877050920318512 | | Hadhami Aouani  Yassine Ben Ayed | | | Emotion recognition, MFCC, ZCR, TEO, HNR, SVM, auto-encoder | |
| **The Name of the Current Solution (Technique/ Method/ Scheme/ Algorithm/ Model/ Tool/ Framework/ ... etc )** | | **The Goal (Objective) of this Solution & What is the problem that need to be solved** | | | **What are the components of it?** | |
| Feature-based SVM Emotion Recognition with Auto-Encoder-Based Feature Dimension Reduction. | | The goal of this solution is to recognize human emotions in speech data using a combination of acoustic features (MFCC, ZCR, TEO, HNR) and machine learning techniques, primarily Support Vector Machines (SVM). The problem that this solution aims to address is the automatic recognition of human emotions in speech data. | | | Feature extraction method**(**MFCC, ZCR, TEO, HNR), Classification model (SVM with Linear, Polynomial & RBF kernel), Feature Dimension Reduction Techniques (Audo Encoder), Ryerson Multimedia Lab (RML) emotion database, Evaluation Metrics | |
| **The Process (Mechanism) of this Work; Means How the Problem has Solved & Advantage & Disadvantage of Each Step in This Process** | | | | | | |
| |  |  |  |  | | --- | --- | --- | --- | |  | **Process Steps** | **Advantage** | **Disadvantage (Limitation)** | | 1 | The acoustic features, including Mel-frequency Cepstral Coefficients (MFCC), Zero Crossing Rate (ZCR), Teager Energy Operator (TEO), and Harmonic-to-Noise Ratio (HNR), are extracted from audio samples. | These features capture important information related to emotions in speech signals, and they are widely used in speech and emotion recognition. | The feature set can be high-dimensional and may contain redundant or irrelevant information. | | 2 | Support Vector Machines (SVM) are employed with different kernels (Linear, Polynomial, and RBF) to classify emotions based on the extracted features. | SVM is a robust classification algorithm, and different kernels allow for flexibility in modeling complex decision boundaries. | SVM performance can be sensitive to kernel and parameter selection. It may not perform optimally with high-dimensional feature sets. | | 3 | Auto-encoders, including Basic AE and Stacked AE, are used to reduce the dimensionality of the feature set by learning a compact representation. | Dimension reduction helps mitigate the curse of dimensionality, reduces computational complexity, and can lead to improved model performance. | The choice of auto-encoder architecture and hyperparameters can impact the quality of the reduced representation. AE training may require careful tuning. | | 4 | Parameters, such as the number of hidden units in auto-encoders, the number of iterations, weight regularization, and SVM kernel parameters, are fine-tuned. | Parameter optimization aims to maximize the system’s accuracy and recognition rates. | Parameter tuning can be time-consuming and may overfit the model to the training data. | | 5 | The system’s performance is evaluated using recognition rates and accuracy for each emotion class. | Metrics provide insights into the system’s ability to recognize specific emotions accurately. | Evaluation metrics do not provide information about misclassification patterns or the system’s generalization to unseen data. | | 6 | The proposed system is compared with other state-of-the-art methods in emotion recognition. | Comparative analysis helps identify the system’s strengths and areas where it outperforms existing approaches. | Results may vary based on the choice of emotion databases and evaluation criteria. | | | | | | | |
| **Major Impact Factors in this Work** | | | | | | |
| |  |  |  |  | | --- | --- | --- | --- | | **Dependent Variable** | **Independent Variable** | **Moderating variable** | **Mediating (Intervening ) variable** | | The dependent variable in this study is the Emotion Classification Outcome, involving the categorization of emotions based on extracted features. The SVM and Auto-Encoder (AE) combined model aims to classify emotions using parameters such as 39 Mel Frequency Cepstral Coefficients (MFCC), Harmonic-to-Noise Ratio (HNR), Zero Crossing Rate (ZCR), and TEO | Theeature variables consist of acoustic and prosodic features extracted from the audio signals. These features include 39 MFCC, HNR, ZCR, and TEO, which collectively form the feature vectors used in the classification algorithm. The SVM and AE act on these independent variables for emotion classification. | The moderating variable in this study is the Feature Dimension Reduction Method, specifically the use of Auto-Encoder (AE). AE serves as a moderating factor in the recognition of emotions by extracting relevant features from the initial parameters, contributing to improved SVM performance. | The mediating variable involves the Learned Feature Representation obtained through the output of the Auto-Encoder. The AE model learns a reduced informative representation of the data, acting as an intermediary step in feature selection before inputting the refined features into the SVM classifier. | | | | | | | |
| |  | | --- | | **Relationship Among the Above 4 Variables in This article** | | The Independent Variable (IV) encompasses acoustic and prosodic features like 39 MFCC, HNR, ZCR, and TEO. These features directly impact the Dependent Variable (DV), representing the emotion classification outcome. The Feature Dimension Reduction Method, serving as a Moderating Variable, employs an Auto-Encoder to compress input features before SVM classification. Simultaneously, the Auto-Encoder’s learned feature embeddings act as the Mediating Variable, optimizing the relationship between the Independent and Dependent Variables. This intricate interplay enhances the effectiveness of emotion recognition in the proposed system. | | | | | | | |
| **Input and Output** | | | **Feature of This Solution** | | | **Contribution & The Value of This Work** |
| |  |  | | --- | --- | | **Input** | **Output** | | Audio samples containing human speech -RML emotion database | Classification and recognition of emotions in audiovisual data. | | | | This solution comprises audiovisual emotion recognition using fused audio features like MFCC, HNR, ZCR, and TEO, enhanced with auto-encoders for feature reduction. SVM is employed for emotion classification. It surpasses existing techniques, particularly in emotion identification from audiovisual data. | | | This work contributes an effective emotion recognition system by combining HNR with traditional audio features and utilizing auto-encoders for feature reduction, yielding improved accuracy. It holds potential for broader applications and larger datasets. |
| **Positive Impact of this Solution in This Project Domain** | | | | **Negative Impact of this Solution in This Project Domain** | | |
| This solution enhances the accuracy of emotion recognition in audio data, which is valuable for various applications like human-computer interaction and sentiment analysis in speech. It provides a foundation for more reliable emotion detection in diverse linguistic contexts, improving user experiences and data-driven decision-making. | | | | One potential negative impact could be the increased computational complexity due to feature extraction and dimension reduction, which might not be suitable for real-time applications with limited resources. | | |
| **Analyse This Work by Critical Thinking** | | | **The Tools That Assessed this Work** | | | **What is the Structure of this Paper** |
| This work introduces an innovative approach to emotion recognition using a combination of acoustic features and auto-encoders for dimension reduction, showing promising results. However, it may lack generalizability to different datasets and requires further evaluation across diverse linguistic and emotional contexts. | | | Scikit-learn, TensorFlow, PyTorch, Matplotlib, Seaborn | | | 1. Abstract 2. Introduction 3. Related work 4. Methods 5. Proposed system 6. Experiments and results 7. Conclusion |
| **Diagram/Flowchart** | | | | | | |
|  | | | | | | |

|  |
| --- |
| **3** |
| **Reference in APA format** | |  | | | | |
| **URL of the Reference** | | **Authors Names and Emails** | | | **Keywords in this Reference** | |
| <https://link>.springer.com/article/10.1007/s10772-018-09572-8 | | S. Lalitha  Shikha Tripathi  Deepa Gupta | | | Arousal, BFCC, Cepstrum , DNN , Emotion detection , Perceptual features , Recognition accuracy , Valence | |
| **The Name of the Current Solution (Technique/ Method/ Scheme/ Algorithm/ Model/ Tool/ Framework/ ... etc )** | | **The Goal (Objective) of this Solution & What is the problem that need to be solved** | | | **What are the components of it?** | |
| Deep neural networks (DNN) along with various perceptual speech features for speech emotion recognition (SER) | | The goal is to develop a Speech Emotion Recognition (SER) system using DNNs and perceptual features to classify and map human emotions in spoken language. The problem is recognizing emotions in speech, both categorically and continuously, using the Berlin corpus as the dataset. | | | Pre-processing module(Involves framing and windowing without filtering the speech signal), Speech feature extraction module(Extracts various speech features from Mel, Bark, and Inverted Mel filter banks, as well as additional features like fundamental frequency, LPC, and functionals), Classification module(DNN), Performance metrics | |
| **The Process (Mechanism) of this Work; Means How the Problem has Solved & Advantage & Disadvantage of Each Step in This Process** | | | | | | |
| |  |  |  |  | | --- | --- | --- | --- | |  | **Process Steps** | **Advantage** | **Disadvantage (Limitation)** | | 1 | Pre-processing includes framing and windowing without any filtering of the speech signal. It prepares the input data for feature extraction. | Keeps the speech signal intact without filtering, which may be important for preserving emotional information. | May not address noise or other signal quality issues. | | 2 | In Speech feature extraction various speech features are extracted from Mel, Bark, and Inverted Mel filter banks, as well as additional features like fundamental frequency, LPC, and functionals. | Provides a rich set of features that capture different aspects of the speech signal related to emotions. | Increases the dimensionality of the feature space, potentially leading to increased computational complexity. | | 3 | A Feed-Forward Back Propagation Network (DNN) is used for emotion classification based on the extracted features. | DNNs are capable of modeling complex, nonlinear relationships in the data, which can improve emotion classification performance. | Training DNNs can be computationally intensive, and they may require a large amount of labeled data for effective training. | | 4 | The system's performance is evaluated using recognition accuracy and confusion matrix in both categorical and continuous emotion dimensions. | Provides quantitative measures of the system’s ability to recognize emotions. | Performance evaluation metrics do not provide insights into the interpretability of the DNN model or its generalization to other datasets. | | | | | | | |
| **Major Impact Factors in this Work** | | | | | | |
| |  |  |  |  | | --- | --- | --- | --- | | **Dependent Variable** | **Independent Variable** | **Moderating variable** | **Mediating (Intervening ) variable** | | This variable refers to the system’s effectiveness in recognizing emotions from speech signals.  **Examples:** Recognition accuracy (%) in both 1-dimensional (categorical) and 2-dimensional (valence and arousal) emotion spaces. | Auditory cues derived from speech signals, including Mel frequency cepstral coefficients (MFCC’s), perceptual linear predictive cepstrum (PLPC), Mel frequency perceptual linear prediction cepstrum (MFPLPC), bark frequency cepstral coefficients (BFCC), revised perceptual linear prediction coefficient’s (RPLP), and inverted Mel frequency cepstral coefficients (IMFCC). | Potential moderating factors could include variations in speech patterns among different individuals or environmental conditions during recording. | **Perceptual Features (MFCC, PLPC, MFPLPC, BFCC, RPLP, IMFCC):** Speech features acting as mediators between the independent variable (input to the system) and the dependent variable (emotion recognition performance). | | | | | | | |
| |  | | --- | | **Relationship Among The Above 4 Variables in This article** | | The relationship among these variables in the article involves the interplay between the chosen perceptual features, the architecture of the DNN, and their combined impact on the system’s ability to recognize emotions from speech signals. The absence of an explicitly mentioned moderating variable suggests a focus on the direct influence of the chosen variables on emotion recognition performance. | | | | | | | |
| **Input and Output** | | | **Feature of This Solution** | | | **Contribution & The Value of This Work** |
| |  |  | | --- | --- | | **Input** | **Output** | | Audio signal containing human speech, specifically speech samples from the Berlin corpus | Prediction or classification of the emotional state conveyed in the input audio signal | | | | This solution combines deep neural networks, such as Feed-Forward Back Propagation Networks, with perceptual speech features derived from Mel and Bark filter banks. It enables the classification of various emotions in speech, including fear, anger, boredom, and more, in both categorical and continuous emotion spaces. The system offers a compact feature vector, speaker independence, and high recognition accuracy, demonstrating its effectiveness for emotion detection in audio signals. The potential for multi-corpus and multimodal applications, as well as insights into addressing imbalanced datasets, adds to its feature set. | | | This work contributes by systematically exploring and identifying significant perceptual features for emotion detection in speech, enabling improved recognition accuracy. The utilization of Feed-Forward Back Propagation Networks offers an effective classification mechanism, particularly when combined with the selected features. It adds value by enhancing the understanding of emotional content in speech, applicable to various domains such as human-computer interaction and sentiment analysis. |
| **Positive Impact of this Solution in This Project Domain** | | | | **Negative Impact of this Solution in This Project Domain** | | |
| In the context of this project, the positive impact of this solution lies in its ability to enhance emotion recognition accuracy in speech, which is crucial for applications like human-computer interaction, sentiment analysis, and affective computing. The identified perceptual features and DNN classification method offer a valuable tool for understanding and extracting emotional content from audio data, ultimately improving the project's performance and usability in emotion-related tasks. | | | | It relies on significant computational resources and specific datasets, limiting its scalability and generalizability to different languages and cultures. Additionally, the feature engineering process may require domain expertise, hindering accessibility for non-experts. | | |
| **Analyse This Work By Critical Thinking** | | | **The Tools That Assessed this Work** | | | **What is the Structure of this Paper** |
| This work offers a promising approach to speech emotion recognition (SER) by combining deep neural networks (DNNs) with selected perceptual speech features. However, while it demonstrates competitive performance, there is room for improvement in feature selection and architecture choice to enhance recognition across diverse datasets and real-world applications. | | | TensorFlow, PyTorch, Keras, NumPy, pandas, Matplotlib, Scikit-learn | | | 1. Abstract 2. Introduction 3. Related work 4. Data source 5. Proposed system 6. Experimental set‑up 7. Experimental results and analysis 8. Performance evaluation 9. Conclusion and outlook |
| **Diagram/Flowchat** | | | | | | |
|  | | | | | | |

|  |
| --- |
| **4** |
| **Reference in APA format** | |  | | | | |
| **URL of the Reference** | | **Authors Names and Emails** | | | **Keywords in this Reference** | |
| <https://ieeexplore>.ieee.org/document/9330527 | | [Bassam Ali Al-Qatab](https://ieeexplore.ieee.org/author/38276950600)  [Mumtaz Begum Mustafa](https://ieeexplore.ieee.org/author/38241272200) | | | Acoustic features, automatic dysarthric speech recognition system, dysarthria, classification algorithms, feature selection methods. | |
| **The Name of the Current Solution (Technique/ Method/ Scheme/ Algorithm/ Model/ Tool/ Framework/ ... etc )** | | **The Goal (Objective) of this Solution & What is the problem that need to be solved** | | | **What are the components of it?** | |
| Automatic speech recognition (ASR) systems | | The goal of this solution is to improve the accuracy of automatic speech recognition (ASR) systems for dysarthric speakers. The problem is that ASR systems are not very accurate when used with dysarthric speakers. | | | Feature selection methods: Interaction Capping (ICAP), Conditional Information Feature Extraction (CIFE), Conditional Mutual Information Maximization (CMIM), Double Input Symmetrical Relevance (DISR), Joint Mutual Information (JMI), Conditional redundancy (Condred) and Relief.  Classification algorithms: Support Vector Machine (SVM), Linear Discriminant Analysis (LDA), Artificial Neural Network (ANN), Classification and Regression Tree (CART), I Bayes (NB), and Random Forest (RF).  Acoustic features : prosody, spectral, cepstral, and voice quality. | |
| **The Process (Mechanism) of this Work; Means How the Problem has Solved & Advantage & Disadvantage of Each Step in This Process** | | | | | | |
| |  |  |  |  | | --- | --- | --- | --- | |  | **Process Steps** | **Advantage** | **Disadvantage (Limitation)** | | 1 | selecting the NEMOURS database, which contains recorded speech from a single dysarthric speaker with varying levels of severity. | The NEMOURS database offers a standardized and consistent dataset for analysis, ensuring the replicability of research. | The limited number of speakers may not fully represent the diversity of dysarthric speech. | | 2 | A wide range of acoustic features is extracted from the speech data, including prosodic, spectral, cepstral, and voice quality features. | Extracting multiple features enables a comprehensive analysis of different aspects of dysarthric speech, potentially capturing valuable information. | The large number of features can make subsequent classification challenging due to dimensionality issues, requiring feature selection. | | 3 | The study employs a feature selection process to reduce the dimensionality of the data. A method based on a formula is used to select a specific number of features. | Feature selection reduces computational complexity, and it can help identify the most relevant features, potentially improving classification accuracy. | The choice of the feature selection method may not be definitive, and the formula used may not be the best fit for all datasets. Different methods could yield different results. | | 4 | Six different classification algorithms, including Support Vector Machine (SVM), Linear Discriminant Analysis (LDA), Artificial Neural Network (ANN), Classification and Regression Tree (CART), I Bayes (NB), and Random Forest (RF), are used to classify the severity of dysarthric speech. | The variety of classification algorithms allows for a comprehensive evaluation, identifying which approach works best for the problem. | The choice of classification algorithms can significantly impact the results, and it may be challenging to determine the best approach. | | 5 | The classification models' performance is evaluated using k-fold cross-validation with k=10, providing a robust estimate of accuracy. | Cross-validation reduces overfitting risks and offers a more reliable assessment of classifier performance. | The choice of the number of folds (k) can affect the results, and the study focused on accuracy as the primary evaluation measure, which may not capture all aspects of performance. | | 6 | Friedman’s m statistics are applied to rank the classifiers and feature selection methods based on their classification accuracy. | Ranking helps identify the most effective combinations of classifiers and feature selection methods, simplifying the selection of the best approach. | Rankings are based on accuracy measures, and any bias in these measures could lead to incorrect rankings. | | | | | | | |
| **Major Impact Factors in this Work** | | | | | | |
| |  |  |  |  | | --- | --- | --- | --- | | **Dependent Variable** | **Independent Variable** | **Moderating variable** | **Mediating (Intervening ) variable** | | **Target Variable – Dysarthric Speech Severity Classification:** This represents the primary outcome variable, categorized into severity levels, serving as the target for the classification. | A. Acoustic Features:  Prosody, Spectral, Cepstral, and Voice Quality Parameters  b. Feature Selection Methods:  Interaction Capping (ICAP), Conditional Information Feature Extraction (CIFE), Conditional Mutual Information Maximization (CMIM), Double Input Symmetrical Relevance (DISR), Joint Mutual Information (JMI), Conditional Redundancy (Condred), and Relief | **Speech Database Size:** This variable moderates the relationship between the independent variables (acoustic features, feature selection methods) and the dependent variable by influencing the effectiveness of the ASR system. | **Feature Selection Methods:** Specifically, the algorithms like ICAP, CIFE, CMIM, DISR, JMI, Condred, and Relief act as mediating variables. They intervene in the relationship between the independent variables (acoustic features) and the dependent variable (Dysarthric Speech Severity Classification). | | | | | | | |
| |  | | --- | | **Relationship Among The Above 4 Variables in This article** | | The acoustic features play a direct role in determining the Dysarthric Speech Severity Classification. However, the impact of these features is shaped and mediated by the chosen feature selection methods. The moderating variable, speech database size, further influences the overall relationship by adjusting the effectiveness of the ASR system in classifying dysarthric speech. | | | | | | | |
| **Input and Output** | | | **Feature of This Solution** | | | **Contribution in This Work** |
| |  |  | | --- | --- | | **Input** | **Output** | | Dysarthric speech | Methodology for improving ASR | | | | This is simply an integration of four acoustic features and seven feature selection methods to design a hybrid one. We can still integrate other feature selection methods which gives us even more better results. | | | Designing hybrid classifier is a good thought, where four acoustic features and seven feature selection methods working together to resolve individual issues. The proposed methodology is a valuable contribution to the field of dysarthric speech recognition. The hybrid classifier has the potential to improve the quality of life for people with dysarthria by improving the accuracy of ASR systems. |
| **Positive Impact of this Solution in This Project Domain** | | | | **Negative Impact of this Solution in This Project Domain** | | |
| Hybrid classifier has the potential to improve the accuracy of ASR systems for dysarthric speakers by integrating individual advantages of acoustic features and feature selection methods.Overall, the proposed methodology is a valuable contribution to the field of dysarthric speech recognition. | | | | The proposed methodology is not computationally efficient and it is trained on a specific dataset that may not be generalizable to other populations of dysarthric speakers. | | |
| **Analyse This Work By Critical Thinking** | | | **The Tools That Assessed this Work** | | | **What is the Structure of this Paper** |
| The proposed methodology is a valuable contribution to the field of dysarthric speech recognition. The authors make a valuable contribution by integrating existing algorithms in a new way to improve the accuracy of ASR systems for dysarthric speakers. | | | openSMILE, MATLAB, Libsvm, Random Forest(RF) | | | 1. Abstract 2. Introduction 3. Related Work 4. Proposed Methodology 5. Experiments and Results 6. Conclusion |
| **Diagram/Flowchart** | | | | | | |
|  | | | | | | |

|  |
| --- |
| **5** |
| **Reference in APA format** | |  | | | | |
| **URL of the Reference** | | **Authors Names and Emails** | | | **Keywords in this Reference** | |
| <https://www>.researchgate.net/publication/321243350\_Detecting\_Stuttering\_Events\_in\_Transcripts\_of\_Children’s\_Speech | | Sadeen Alharbi  Madina Hasan  Anthony J H Simons  Shelagh Brumfitt &  Phil Green | | | Stuttering event detection, Speech disorder, human-computer interaction, CRF, HELM | |
| **The Name of the Current Solution (Technique/ Method/ Scheme/ Algorithm/ Model/ Tool/ Framework/ ... etc )** | | **The Goal (Objective) of this Solution & What is the problem that need to be solved** | | | **What are the components of it?** | |
| Machine learning approaches, specifically the “Hidden Event Language Model (HELM)” and “Conditional Random Fields (CRF). | | The go”l of “he solution is to automatically detect and classify stuttering events in transcripts of children’s speech. This addresses the problem of accurately identifying stuttering in speech, which can impact a child’s development if left untreated. The solution aims to provide a faster and more reliable method for early diagnosis and intervention. | | | Data Transcription and Annotation, Data Normalization and Feature Extraction ( extracting word-level features, including n-grams and post-words), Classification Approaches (HELM & CRF), Data Augmentation, Metrics( precision, recall, F1 score, and accuracy). | |
| **The Process (Mechanism) of this Work; Means How the problem has Solved & Advantage & Disadvantage of Each Step in This Process** | | | | | | |
| This process combines data collection, preprocessing, machine learning models, data augmentation, and evaluation to detect and classify stuttering events in children’s speech transcripts.   |  |  |  |  | | --- | --- | --- | --- | |  | **Process Steps** | **Advantage** | **Disadvantage (Limitation)** | | 1 | The work starts by collecting orthographic transcripts of children’s speech, which serves as the input data. Annotators label different types of stuttering events within the text using a specific annotation approach. | This step provides the necessary data to train and test the classification models, making it possible to analyze stuttering in children’s speech. | Collecting and annotating the data can be time-consuming and may rely on the availability of such data. | | 2 | Text normalization is performed to prepare the data for analysis. This step includes transforming text entities (e.g., dates, numbers, and times) into words and extracting word-level features, including n-grams and post-words. | Data preprocessing ensures that the data is in a suitable format for machine learning analysis, making it easier to extract relevant features. | The choice of features and the preprocessing process may impact the quality of the data and the results. | | 3 | The work employs two machine learning approaches, HELM and CRF, to classify and detect stuttering events within the transcripts based on the extracted features. | Machine learning models can automatically detect stuttering events in transcripts, potentially reducing the subjectivity associated with manual assessment. | The performance of the models depends on the quality and quantity of training data and the choice of features. | | 4 | To improve the models’ ability to detect stuttering events, the work generates additional sentences with stuttering patterns, supplementing the original training data. | Data augmentation increases the diversity and volume of the training data, enhancing the models’ performance, especially for rare stuttering events. | The quality of augmented data may vary, and it could introduce noise into the training data. | | 5 | The solution employs standard evaluation metrics such as precision, recall, F1 score, and accuracy to assess the performance of the models in detecting stuttering events in the transcripts. | Evaluation metrics provide a quantitative measure of the solution’s effectiveness in identifying stuttering, enabling comparisons and improvements. | The choice of evaluation metrics may not capture all aspects of model performance, and some stuttering events may remain challenging to detect. | | | | | | | |
| **Major Impact Factors in this Work** | | | | | | |
| |  |  |  |  | | --- | --- | --- | --- | | **Dependent Variable** | **Independent Variable** | **Moderating variable** | **Mediating (Intervening ) variable** | | Stuttering Event Detection Accuracy | Machine learning approaches (HELM and CRF).  Data augmentation techniques. | Training Data Availability. | Text Normalization features  Word-level features  In the context of the classification approaches:  **1.HELM (Hidden Event Language Model):**  Probability of stuttering events at the end of each observed word given its context.  **2.CRFs (Conditional Random Fields):**  Sequence labelling and segmentation of stuttering events.  Estimation and optimization of the posterior probability of the label sequence given a sequence of features. | | | | | | | |
| |  | | --- | | **Relationship Among the Above 4 Variables in This article** | | The chosen machine learning approaches and data augmentation techniques influence the accuracy of detecting stuttering events. However, the presence or absence of sufficient training data moderates this relationship. The process of text normalization and the extracted word-level features serves as a mediating mechanism, influencing how the machine learning approaches impact the accuracy of stuttering event detection. | | | | | | | |
| **Input and Output** | | | **Feature of This Solution** | | | **Contribution & The Value of This Work** |
| |  |  | | --- | --- | | **Input** | **Output** | | orthographic transcripts of children’s speech. | Diagnosis of stuttering events in children’s speech transcripts. | | | | The solution's key feature is its use of machine learning approaches (HELM and CRF) to detect and classify stuttering events in children’s speech transcripts, along with data augmentation to improve performance for rare events. | | | The work contributes to the field of stuttering event detection by introducing machine learning techniques and data augmentation to enhance accuracy and addresses the scarcity of training data. Its value lies in improving the assessment of stuttering in children, aiding early intervention, and expanding the annotated speech data available for research. |
| **Positive Impact of this Solution in This Project Domain** | | | | **Negative Impact of this Solution in This Project Domain** | | |
| The positive impact of this solution in the project domain is more accurate and efficient stuttering event detection in children’s speech, facilitating early intervention and research in speech disorders. | | | | The negative impact of this solution in the project domain might be the reliance on generated data for training, which could introduce noise and inaccuracies in the models, potentially affecting the precision of stuttering event detection. | | |
| **Analyse This Work By Critical Thinking** | | | **The Tools That Assessed this Work** | | | **What is the Structure of this Paper** |
| This work leverages machine learning models like HELM and CRF for automated stuttering event detection in children’s speech transcripts, potentially aiding early intervention in speech disorders. However, challenges persist in detecting less frequent stuttering events and ethical aspects related to data augmentation. | | | Hidden Event Language Model (HELM), Conditional Random Fields (CRF), Automatic Speech Recognition (ASR), and the SRILM toolkit for data augmentation | | | 1. Abstract 2. Introduction 3. Data Transcription and Annotation 4. Data Normalisation and Features Extraction 5. Classification Approaches 6. Data Augmentation 7. Metrics 8. Experiments 9. Conclusions and Future Work |
| **Diagram/Flowchart** | | | | | | |
|  | | | | | | |

|  |
| --- |
| **6** |
| **Reference in APA format** | |  | | | | | |
| **URL of the Reference** | | **Authors Names and Emails** | | | | **Keywords in this Reference** | |
| <https://www.sciencedirect.com/science/article/pii/S1877050920318512> | | Hadhami Aouani  Yassine Ben Ayed | | | | Emotion recognition, Mel-frequency Cepstral Coefficient, Zero Crossing Rate, Teager Energy Operator, Harmonic to Noise Ratio, Support Vector Machine, Auto-encoder | |
| **The Name of the Current Solution (Technique/ Method/ Scheme/ Algorithm/ Model/ Tool/ Framework/ ... etc )** | | **The Goal (Objective) of this Solution & What is the problem that need to be solved** | | | | **What are the components of it?** | |
| Support Vector Machine (SVM) | | The objective is to create a robust emotion recognition system using speech signals. The challenge is to automatically identify and classify human emotions. The solution combines features like MFCC, HNR, ZCR, and TEO, utilizing SVM for classification. Auto-encoders are introduced for feature dimension reduction to enhance SVM performance. | | | | An emotion recognition system based on speech signals in two-stage approach, namely feature extraction and classification engine. | |
| **The Process (Mechanism) of this Work; Means How the Problem has Solved & Advantage & Disadvantage of Each Step in This Process** | | | | | | | |
| The paper proposes a new framework for speech emotion recognition using one-dimensional deep convolutional neural networks (CNNs) with the combination of five different audio features as input data. The authors evaluate their model on three public datasets, RAVDESS, IEMOCAP, and EMO-DB, and achieve state-of-the-art results on all three datasets.   |  |  |  |  | | --- | --- | --- | --- | |  | **Process Steps** | **Advantage** | **Disadvantage (Limitation)** | | **1** | 1. Extract five different features from a sound file:    * Mel-frequency cepstral coefficients (MFCCs)    * Spectral centroid    * Spectral roll-off    * Zero-crossing rate    * Energy 2. And Feed the extracted features into a one-dimensional convolutional neural network (CNN) | * The use of deep CNNs, which are able to learn complex patterns in the data without the need for handcrafted feature engineering. | While deep learning models have achieved state-of-the-art results on some public datasets, their accuracy in real-world conditions can be lower. | | **2** | 1. Train the CNN to predict the emotion of the speaker and Evaluate the trained CNN on a held-out test set | The combination of multiple audio features, which provides more information to the model and helps to improve its performance | Deep learning models can be sensitive to noise and interference in the speech signal. This can degrade their performance, especially in noisy environments. | | | | | | | | |
| **Major Impact Factors in this Work** | | | | | | | |
| |  |  |  |  | | --- | --- | --- | --- | | **Dependent Variable** | **Independent Variable** | **Moderating variable** | **Mediating (Intervening ) variable** | | The emotion | Features used for emotion recognition like MFCC, HNR, ZCR, TEO | Auto-encoder | SVM Classifier | | | | | | | | |
| |  | | --- | | **Relationship Among The Above 4 Variables in This article** | | The problem is accurately identifying emotions from speech. Features like MFCC, HNR, ZCR, and TEO are independent variables. Emotion identification is the dependent variable. Auto-encoder moderates by reducing feature dimensionality, and SVM classifies emotions, serving as a mediating variable. Together, they create a system for enhanced speech emotion recognition. | | | | | | | | |
| **Input and Output** | | | **Feature of This Solution** | | | | **Contribution & The Value of this Work** |
| |  |  | | --- | --- | | **Input** | **Output** | | a raw audio file, or features extracted from the audio signal | A prediction of the speaker’s emotion. | | | | It uses a combination of five different spectral representations of the same sound file as input to a deep learning model. This combination of features allows the model to better capture the nuances of human speech and emotion. | | | | A new framework for speech emotion recognition using deep learning. The authors combine five audio features as input to a one-dimensional deep CNN, and their model outperforms the state-of-the-art on two datasets. |
| **Positive Impact of this Solution in This Project Domain** | | | | | **Negative Impact of this Solution in This Project Domain** | | |
| It has the potential to improve the user experience, educational outcomes, customer service quality, healthcare delivery, and marketing effectiveness. | | | | | The Speech emotion recognition with deep learning can be biased, invade privacy, and be used for manipulation or security breaches. | | |
| **Analyse This Work By Critical Thinking** | | | | **The Tools That Assessed this Work** | | | **What is the Structure of this Paper** |
| This work is good and promising, but it is important to be aware of its limitations and potential negative impacts. It is important to use speech emotion recognition responsibly and to mitigate the potential risks. | | | | Support Vector Machine | | | Abstract   1. Introduction 2. Related work 3. Methods 4. Experiments and results 5. Conclusion |
| **Diagram/Flowchart** | | | | | | | |
|  | | | | | | | |

|  |
| --- |
| **7** |
| **Reference in APA format** | |  | |
| **URL of the Reference** | | **Authors Names and Emails** | **Keywords in this Reference** |
| <https://ijrpr.com/uploads/V3ISSUE5/IJRPR4210.pdf> | | Husbaan I. Attar, Nilesh K. Kadole, Omkar G. Karanjekar, Devang R. Nagarkar, Prof. Sujeet | Speech Emotion Recognition (SER), Convolutional Neural Networks (CNN),Emotion Classification,Human-Machine Interaction Deep Neural Networks (DNN) |
| **The Name of the Current Solution (Technique/ Method/ Scheme/ Algorithm/ Model/ Tool/ Framework/ ... etc )** | | **The Goal (Objective) of this Solution & What is the problem that need to be solved** | **What are the components of it?** |
| Convolutional Neural Networks | | The goal of this solution is to advance affective computing, facilitating effective human-computer interaction by proposing a Real-Time Speech Emotion Recognition System. It aims to overcome this limitation by developing a system that recognizes emotions in real-time from continuous speech. | It consists of Voice Activity Detection to identify speech segments, Speech Segmentation for meaningful division, Signal Pre-Processing for conditioning audio, Feature Extraction extracting relevant speech features, Emotion Classification utilizing machine learning, and Statistics Analysis of Emotion Frequency for insight. |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **The Process (mechanism) of this Work; Means How the Problem has Solved & Advantage & Disadvantage of Each Step in This Process** | | | | |
| |  |  |  |  | | --- | --- | --- | --- | |  | **Process Steps** | **Advantage** | **Disadvantage (Limitation)** | | **1** | It involves gathering speech data, particularly from the RAVDESS dataset, and pre-processing it by extracting features like Mel Frequency Cepstral Coefficients (MFCCs) using LibROSA. | 1. The proposed speech emotion recognition system demonstrates a high accuracy rate, reaching 90% in experiments. | * The system’s effectiveness may vary depending on the complexity and nuances of different emotional states. | | **2** | 1. A Convolutional Neural Network (CNN) architecture is employed for speech emotion recognition. | 1. The system’s application in online learning environments proves beneficial by efficiently recognizing students’ responses to the course material. | * While the system performs well with pre-recorded datasets, its reliance on such data may pose challenges when faced with diverse and dynamic real-world scenarios. | | **3** | 1. The trained model is rigorously evaluated using both pre-recorded datasets and real-time recordings featuring various emotion categories. |  |  | | | | | |
| **Major Impact Factors in this Work** | | | | |
| |  |  |  |  | | --- | --- | --- | --- | | **Dependent Variable** | **Independent Variable** | **Moderating variable** | **Mediating (Intervening ) variable** | | Stuttering Severity | Automation of Stuttering Recognition | Age | Workload of Speech Language Pathologists | | | | | |
| |  | | --- | | **Relationship Among The Above 4 Variables in This article** | | This study explores the connection between stuttering severity and the automation of stuttering recognition , with age serving as a moderating factor. The workload of Speech Language Pathologists acts as a mediating variable, influencing the relationship between automation and stuttering severity. The study aims to discern how automated recognition, moderated by age and mediated by workload, correlates with the severity of stuttering. | | | | | |
| **Input and Output** | **Feature of This Solution** | | | **Contribution in This Work** |
| |  |  | | --- | --- | | **Input** | **Output** | | The input is speech | The output is emotion | | These features collectively contribute to a comprehensive representation of speech signals, enabling the automated system to effectively recognize and quantify the severity of stuttering. The system’s distinctive Statistics Analysis of Emotion Frequency provides insights into emotional prevalence. | | | * The automatic recognition of stuttering severity is a significant advancement in clinical practice.This system ensures objective measurements, facilitates early interventions, contributes to research endeavours, enhances patient experiences, and promotes public awareness. Moreover, it offers the prospect of cost-efficient healthcare practices through streamlined assessments. |
| **Positive Impact of this Solution in This Project Domain** | | | **Negative Impact of this Solution in This Project Domain** | |
| It introduces a real-time emotion recognition system for continuous speech, elevating human-computer interaction with advanced features like efficient Voice Activity Detection and comprehensive Speech Segmentation. The use of a Convolutional Neural Network (CNN) for Emotion Classification ensures accurate and responsive results, enhancing the user experience with a nuanced and personalized touch. | | | The complexity of implementing advanced features, particularly the CNN, may require specialized skills and resources, potentially limiting accessibility for some organizations. Additionally, the system’s performance could be constrained by the quality and diversity of the training data, posing a risk of reduced accuracy, especially for less-represented emotional states. | |
| **Analyze This Work By Critical Thinking** | | **The Tools That Assessed this Work** | | **What is the Structure of this Paper** |
| This approach has the potential to enhance user experience by providing nuanced and personalized interactions. However, the complexity of implementation, reliance on representative training data, and ethical considerations pose challenges. The system’s success depends on addressing these concerns and staying abreast of evolving technological landscapes and competing solutions. | | A combination of statistical analysis tools, machine learning frameworks like Tensor Flow or PyTorch, and speech processing libraries such as LibROSA. | | Abstract   1. Introduction 2. Problem Statement 3. Existing system 4. Existing system algorithm 5. System Implementation 6. Implementation Diagram   Conclusion |
| **Diagram/Flowchart** | | | | |
|  | | | | |

|  |
| --- |
| **8** |
| **Reference in APA format** | |  | | | | | |
| **URL of the Reference** | | **Authors Names and Emails** | | | | **Keywords in this Reference** | |
| <https://arxiv.org/ftp/arxiv/papers/2007/2007.08003.pdf> | | Dr. Mrs. Gresha Bhatia, Binoy Saha, Mansi Khamkar, Ashish Chandwani , Reshma Khot | | | | Stutter diagnosis, Stuttering therapy, Stutter measurement, Speech dysfluency, Mel-frequency Cepstral Coefficients (MFCC), CNN, Gated Recurrent Units (GRU), Support Vector Machine (SVM) | |
| **The Name of the Current Solution (Technique/ Method/ Scheme/ Algorithm/ Model/ Tool/ Framework/ ... etc )** | | **The Goal (Objective) of this Solution & What is the problem that need to be solved** | | | | **What are the components of it?** | |
| The proposed technique is Support Vector Machine (SVM) and Gated Recurrent CNN (GRCNN) Models | | The main objective is to improve a person’s speech fluency by accurately diagnosing stutter and suggesting appropriate training exercises for practice. | | | | The components are stutter assessment, therapy suggestion, Gated Recurrent CNN (GRCNN) models, SVM model, and a mobile application. | |
| **The Process (Mechanism) of this Work; Means How the Problem has Solved & Advantage & Disadvantage of Each Step in This Process** | | | | | | | |
| |  |  |  |  | | --- | --- | --- | --- | |  | **Process Steps** | **Advantage** | **Disadvantage (Limitation)** | |  | Stutter Assessment involves dataset preparation, data labeling, and feature extraction using MFCC features. | The Stutter Diagnosis and Therapy System Based on Deep Learning provides quantitative analysis of stuttering, allowing for accurate diagnosis and assessment of the severity and type of stutter. | The lack of a well-structured dataset for stuttered speech poses a challenge in training the models and may result in false positives due to background noise | |  | Separate GRCNN models are trained for detecting prolongation and repetition in speech audio. | The system recommends appropriate speech therapies based on the patient’s performance and improvement, providing personalized treatment options. |  | |  | An SVM model with a polynomial kernel is trained to recommend suitable therapies |  |  | | | | | | | | |
| **Major Impact Factors in this Work** | | | | | | | |
| |  |  |  |  | | --- | --- | --- | --- | | **Dependent Variable** | **Independent Variable** | **Moderating variable** | **Mediating (Intervening ) variable** | | Effectiveness of speech therapies | Stutter descriptors | Baseline speech proficiency | correlation between stutter descriptors and effectiveness of speech therapies | | | | | | | | |
| |  | | --- | | **Relationship Among The Above 4 Variables in This article** | | This study investigates how stutter descriptors (independent variable) impact the effectiveness of speech therapies (dependent variable). Baseline proficiency moderates this relationship, while the correlation between stutter descriptors and therapy effectiveness serves as a mediating factor. The goal is to discern the complex dynamics and provide personalized therapy recommendations for individuals with stuttering. | | | | | | | | |
| **Input and Output** | | | **Feature of This Solution** | | | | **Contribution & The Value of This Work** |
| |  |  | | --- | --- | | **Input** | **Output** | | Recorded speech audio | Quantitative analysis of the type and severity of stuttered disfluencies in speech and recommendations | | | | The various features in the solution are Stutter assessment using Mel-frequency Cepstral Coefficients (MFCC) features, therapy suggestion based on patient performance, separate Gated Recurrent CNN (GRCNN) models for detecting prolongation and repetition, an SVM model for recommending suitable therapies, and integration into a mobile application for personalized speech therapy. | | | | It based on Deep Learning uses Mel-frequency Cepstral Coefficients (MFCC) features for stutter assessment and recommends appropriate therapies based on patient performance. The system’s value lies in automating tasks, improving accuracy, and providing an affordable solution for people who stutter. |
| **Positive Impact of this Solution in This Project Domain** | | | | | **Negative Impact of this Solution in This Project Domain** | | |
| By automating stutter assessment and recommending personalized therapies, the system offers an affordable and accessible solution. This technological intervention addresses challenges such as the high cost of private speech therapy and the need for customized treatment plans | | | | | the implementation of technology in stutter therapy raises concerns. Affordability and accessibility issues may persist, especially in regions with limited technological infrastructure. Privacy concerns also emerge as continuous monitoring and analysis of speech patterns become integral to therapy. | | |
| **Analyse This Work By Critical Thinking** | | | | **The Tools That Assessed this Work** | | | **What is the Structure of this Paper** |
| This work would be to include a larger and more diverse dataset of stuttered speech recordings to improve the accuracy and generalizability of the models. It provided a highly innovative and valuable in addressing the need for technology in speech therapy and providing an affordable and accessible solution for people who stutter. | | | | The tools used to assess this work include Python and its libraries (such as PyDub and Librosa) for audio processing and feature extraction, scikit-learn for training the SVM model, Keras with TensorFlow backend for developing and training the GRCNN models, and various deep learning models such as CNN and RNN. | | | Abstract   1. Introduction 2. Previous Work 3. Our Approach 4. Evaluation 5. Key Findings 6. Results 7. Conclusion 8. Future Scope |
| **Diagram/Flowchart** | | | | | | | |
|  | | | | | | | |

|  |
| --- |
| **9** |
| **Reference in APA format** | |  | | | | | |
| **URL of the Reference** | | **Authors Names and Emails** | | | | **Keywords in this Reference** | |
| <https://scholar.google.com/scholar?start=30&q=stuttering+recognition+machine+learning&hl=en&as_sdt=0,5#d=gs_qabs&t=1700053247762&u=%23p%3D6odTPCRxfVIJ> | | Seema Barda | | | | Stuttering, Speech dysfluency, Automatic Speech Recognition (ASR), Stuttering severity, Mel frequency Cepstral Coefficients (MFCC), Support Vector Machine (SVM) | |
| **The Name of the Current Solution (Technique/ Method/ Scheme/ Algorithm/ Model/ Tool/ Framework/ ... etc )** | | **The Goal (Objective) of this Solution & What is the problem that need to be solved** | | | | **What are the components of it?** | |
| It uses Automatic Speech Recognition (ASR), Mel frequency Cepstral Coefficients (MFCC) for feature extraction, and Support Vector Machine (SVM) for feature classification in the context of analyzing stuttering severity. | | Theeaturon aims to automate the assessment of stuttering severity by proposing an approach that uses Automatic Speech Recognition (ASR), focusing on features like Mel frequency Cepstral Coefficients (MFCC) and employing Support Vector Machine (SVM) for classification. | | | | The solution involves using Mel frequency Cepstral Coefficients (MFCC) for feature extraction and Support Vector Machine (SVM) for classification to automate the assessment of stuttering severity. | |
| **The Process (Mechanism) of this Work; Means How the Problem has Solved & Advantage & Disadvantage of Each Step in This Process** | | | | | | | |
| |  |  |  |  | | --- | --- | --- | --- | |  | **Process Steps** | **Advantage** | **Disadvantage (Limitation)** | | **1** | The system begins by pre-processing the input speech signal, including noise reduction and filtering. It then uses Voice Activity Detection to identify speech segments | The system enables real-time recognition of emotions in continuous speech, facilitating immediate and responsive interactions in human-computer interfaces. | Developing and implementing a real-time speech emotion recognition system, especially one based on advanced techniques like CNNs, may require specialized expertise and resources. | | **2** | Extract crucial speech features like MFCCs and spectral attributes, and employ a trained machine learning model, potentially a Convolutional Neural Network (CNN), for efficient emotion classification. | By understanding and classifying emotional states, the system can deliver personalized responses, enhancing user experience and engagement | The system may struggle with recognizing emotions not well-represented in the training set. | | **3** | A statistical analysis on recognized emotions, offering quantitative insights into the prevalence and distribution of different emotional states in the continuous speech data. |  |  | | | | | | | | |
| **Major Impact Factors in this Work** | | | | | | | |
| |  |  |  |  | | --- | --- | --- | --- | | **Dependent Variable** | **Independent Variable** | **Moderating variable** | **Mediating (Intervening ) variable** | | Stuttering Severity | Automated Speech Recognition System | Speech-language pathologists’ expertise level | Efficiency of the Automated Speech Recognition System | | | | | | | | |
| |  | | --- | | **Relationship Among The Above 4 Variables in This article** | | The independent variable is the “Automated Speech Recognition System,” which directly influences the dependent variable, the “Stuttering Severity Assessment.” The efficiency of the recognition system acts as a mediating factor, impacting the accuracy of stuttering assessment. Concurrently, the expertise of “Speech-language pathologists” moderates this relationship, influencing how the system’s effectiveness varies based on the pathologists’ proficiency. | | | | | | | | |
| **Input and Output** | | | **Feature of This Solution** | | | | **Contribution & The Value of This Work** |
| |  |  | | --- | --- | | **Input** | **Output** | | Speech | Severity | | | | This solution utilizes Mel frequency cepstral coefficients (MFCC) for efficient feature extraction and employs Support Vector Machine (SVM) for classification. By analyzing spectral and temporal features, the system distinguishes between normal and stuttered speech, contributing to speech pathology by automating severity assessment. | | | | This work significantly contributes to the field of speech pathology by introducing an automated solution for assessing the severity of stuttering. By leveraging advanced techniques like MFCC and SVM, the system provides an efficient and accurate method, reducing the time-consuming manual efforts. The automation not only enhances the speed of assessment but also allows professionals to focus more on therapeutic strategies. |
| **Positive Impact of this Solution in This Project Domain** | | | | | **Negative Impact of this Solution in This Project Domain** | | |
| The automated stuttering severity assessment system offers significant positive impacts by streamlining diagnostic processes, reducing evaluation time, and enabling speech-language pathologists to concentrate on effective therapeutic interventions, ultimately enhancing patient outcomes and treatment quality. | | | | | There are challenges in terms of accuracy and reliability, as it heavily relies on speech recognition technology. The reliance on technology may limit the system’s effectiveness in capturing the nuanced and subjective aspects of stuttering. | | |
| **Analyse This Work By Critical Thinking** | | | | **The Tools That Assessed this Work** | | | **What is the Structure of this Paper** |
| It should focus on addressing challenges and assessing the real-world implications of relying solely on automated systems for a nuanced condition like stuttering, ensuring it complements rather than replaces human expertise in therapeutic interventions | | | | Mel Frequency Cepstral Coefficients (MFCC) for feature extraction and Support Vector Machine (SVM) for classification. | | | * 1. Abstract   2. Introduction   3. Challenges   4. Types of Speech Recognition Techniques   5. Methodology   6. Conclusion |
| **Diagram/Flowchart** | | | | | | | |
|  | | | | | | | |

|  |
| --- |
| **10** |
| **Reference in APA format** | |  | | | | | |
| **URL of the Reference** | | **Authors Names and Emails** | | | | **Keywords in this Reference** | |
| <https://www.sciencedirect.com/science/article/pii/S2665917423002490> | | M. Mahendran  R. Visalakshi  S. Balaji | | | | Dysarthia, Speech detection, CNN, MFCC Feature extraction | |
| **The Name of the Current Solution (Technique/ Method/ Scheme/ Algorithm/ Model/ Tool/ Framework/ ... etc )** | | **The Goal (Objective) of this Solution & What is the problem that need to be solved** | | | | **What are the components of it?** | |
| The paper proposes a CNN-based model that uses various speech features to detect dysarthria in patients | | The goal of the proposed solution is to detect dysarthria in patients using a CNN-based model that analyses various speech features. The problem that needs to be solved is the difficulty in diagnosing dysarthria, which is speech impairment caused by various underlying conditions. | | | | A Convolutional Neural Network-based model to detect dysarthria in patients. The model analyses various speech features such as zero crossing rates, MFCCs, spectral centroids, and spectral roll off. The TORGO speech signal database is used for training and testing the model. | |
| **The Process (Mechanism) of this Work; Means How the Problem has Solved & Advantage & Disadvantage of Each Step in This Process** | | | | | | | |
| |  |  |  |  | | --- | --- | --- | --- | |  | **Process Steps** | **Advantage** | **Disadvantage (Limitation)** | | **1** | Pre-processing the data, this includes silence removal and noise reduction. | The proposed solution is generalizable across languages, making it useful for detecting dysarthria in patients speaking different languages. | It requires a large amount of data for training the model, which may not always be available. | | **2** | Feature extraction using various speech features such as zero crossing rates, MFCCs, spectral centroids, and spectral roll off. | The end-to-end framework optimizes feature extraction, distance matrix computation, and classification, making the model efficient and effective | The model's accuracy may be affected by variations in speech patterns due to factors such as age, gender, and accent. | | **3** | Training the CNN model using the TORGO speech signal database. | It surpasses state-of-the-art CNN-based systems, according to experimental results on two dysarthric speech datasets | Theeaturon may not be suitable for detecting dysarthria in patients with severe speech impairments, as the model may not be able to capture the unique characteristics of their speech. | | **4** | Testing the model on the same database to evaluate its accuracy in detecting dysarthria. |  |  | | **5** | Using the trained model to detect dysarthria in patients by analysing their speech features. |  |  | | | | | | | | |
| **Major Impact Factors in this Work** | | | | | | | |
| |  |  |  |  | | --- | --- | --- | --- | | **Dependent Variable** | **Independent Variable** | **Moderating variable** | **Mediating (Intervening ) variable** | | Dysarthria Diagnosis | Severity level of dysarthria | Age | Speech features | | | | | | | | |
| |  | | --- | | **Relationship Among The Above 4 Variables in This article** | | This study examines the link between dysarthria severity and the accuracy of diagnosis using a Convolutional Neural Network . Speech features like zero crossing rates and MFCCs serve as mediating variables, explaining how dysarthria severity affects diagnosis accuracy. Age is considered as a potential moderating variable, suggesting its influence on this relationship. Overall, the study showcases the CNN’s effectiveness in early dysarthria diagnosis, considering severity levels and potential moderating factors. | | | | | | | | |
| **Input and Output** | | | **Feature of This Solution** | | | | **Contribution & The Value of This Work** |
| |  |  | | --- | --- | | **Input** | **Output** | | An audio sample of speech | A binary classification label, indicating normal speech or dysarthria speech | | | | The proposed solution uses a Convolutional Neural Network-based model for dysarthria detection, which has an accuracy score of 93.87%. The model uses various speech features such as zero crossing rates, MFCCs, spectral centroids, and spectral roll off for the analysis of speech signals | | | | The contribution of this work is the development of a CNN-based model for dysarthria detection, which has a high accuracy score and uses various speech features for analysis. The value of this work lies in its potential to improve the quality of life for individuals with dysarthria by enabling early detection and intervention. |
| **Positive Impact of this Solution in This Project Domain** | | | | | **Negative Impact of this Solution in This Project Domain** | | |
| It can aid in early detection and better management of dysarthria, leading to improved quality of life for individuals with the impairment. Additionally, the proposed model shows promising results in detecting dysarthria with high accuracy, which can potentially reduce the need for invasive diagnostic procedures. | | | | | It may not be accessible to individuals who do not have access to the necessary technology or resources. Additionally, the model’s effectiveness may be limited to the specific dysarthric speech datasets used in the study, and further research is needed to validate its effectiveness on a larger scale. | | |
| **Analyse This Work By Critical Thinking** | | | | **The Tools That Assessed this Work** | | | **What is the Structure of this Paper** |
| This work is a significant step towards improving the quality of life for individuals with dysarthria, but further research is needed to validate the proposed model's effectiveness. | | | | The tools used to assess this work include a pairwise distance-based CNN, feature extraction, distance matrix computation, and classification. The model was evaluated on two dysarthria speech datasets, and the results showed that it outperformed state-of-the-art CNN-based systems. | | | 1. Introduction  2. Related Work  3. Materials and Methods  3.1. Data Collection  3.2. Data Preprocessing  3.3. Feature Extraction  3.4. Convolutional Neural Network  3.5. Model Training and Evaluation  4. Results and Discussion  5. Conclusion and Future Work  6. References |
| **Diagram/Flowchart** | | | | | | | |
|  | | | | | | | |

|  |
| --- |
| **11** |
| **Reference in APA format** | |  | | | | | |
| **URL of the Reference** | | **Authors Names and Emails** | | | | **Keywords in this Reference** | |
| <https://www>.mdpi.com/1424-8220/20/1/183 | | Mustaqeem  Soonil Kwon | | | | Artificial Intelligence, emotion recognition,convolutional neural networks (CNN), noise removal, spectrogram, signals enhancement | |
| **The Name of the Current Solution (Technique/ Method/ Scheme/ Algorithm/ Model/ Tool/ Framework/ ... etc )** | | **The Goal (Objective) of this Solution & What is the problem that need to be solved** | | | | **What are the components of it?** | |
| An artificial intelligence-assisted deep stride convolutional neural network (DSCNN) | | Increasing the accuracy of speech emotion recognition and reducing the computational complexity of the model. | | | | The system includes three components:  1.Audio signal preprocessing for spectrogram generation.  2. Deep stride convolutional neural network (DSCNN) architecture for extracting deep features from the spectrograms and classify the emotions.  3. post-processing, involves in decoding the predicted emotion labels and generating the final output. | |
| **The Process (Mechanism) of this Work; Means How the Problem has Solved & Advantage & Disadvantage of Each Step in This Process** | | | | | | | |
| |  |  |  |  | | --- | --- | --- | --- | |  | **Process Steps** | **Advantage** | **Disadvantage (Limitation)** | | **1** | Data is pre-processed to remove noisy and silent portions. IEMOCAP and RAVDES datasets are used. | The model has a training time of only 14 minutes, demonstrating its computational efficiency compared to other state-of-the-art CNN models | Computational cost is more to train and run. | | **2** | Spectrograms are created from the speech data and are fed into a CNN model to extract features. | The model has a smaller model size compared to other state-of-the-art CNN models, making it computationally simpler | The proposed model may require significant computational resources for training on larger datasets. | | **3** | Features are used to classify emotion of the speaker. | The proposed CNN model improves the overall prediction accuracy, indicating its robustness | The model may not perform well on noisy data. | | | | | | | | |
| **Major Impact Factors in this Work** | | | | | | | |
| |  |  |  |  | | --- | --- | --- | --- | | **Dependent Variable** | **Independent Variable** | **Moderating variable** | **Mediating (Intervening ) variable** | | Speech emotion recognition accuracy | Deep Stride Convolutional Neural Network | Specific features learned in the DSCNN | Characteristics of datasets | | | | | | | | |
| |  | | --- | | **Relationship Among The Above 4 Variables in This article** | | The accuracy of speech emotion recognition hinges on the utilization of a Deep Stride Convolutional Neural Network (DSCNN), with the specific features learned within the network directly influencing this accuracy. Moreover, the characteristics of datasets serve as a crucial moderating factor, shaping how the DSCNN and its learned features interact to affect the overall accuracy of speech emotion recognition. | | | | | | | | |
| **Input and Output** | | | **Feature of This Solution** | | | | **Contribution & The Value of This Work** |
| |  |  | | --- | --- | | **Input** | **Output** | | Audio signals | Speaker emotion | | | | The system features audio signal preprocessing, a DSCNN architecture for feature extraction, and emotion classification, followed by post-processing for generating the final output. | | | | DSCNN model and the adaptive threshold-based pre-processing of the speech signal improved the accuracy of speech emotion. |
| **Positive Impact of this Solution in This Project Domain** | | | | | **Negative Impact of this Solution in This Project Domain** | | |
| Accuracy, identify user emotional state for appropriate response and selection of the classifiers with outlier detection enhances the speech emotion recognition process. | | | | | Oversimplification of human emotions, misinterpretation of emotions. | | |
| **Analyse This Work By Critical Thinking** | | | | **The Tools That Assessed this Work** | | | **What is the Structure of this Paper** |
| The proposed method first enhances the audio signal using a novel method based on Convolutional neural networks, and then classifies the emotions using SVM. The method is novel and effective and the experimental results are convincing. | | | | Convolutional neural network (CNN), Deep Neural Network (DNN) | | | Abstract   1. Introduction 2. Related Work 3. Proposed Methodology 4. Experiments and Results 5. Conclusion |
| **Diagram/Flowchart** | | | | | | | |
|  | | | | | | | |

|  |
| --- |
| **12** |
| **Reference in APA format** | |  | | | | | |
| **URL of the Reference** | | **Authors Names and Emails** | | | | **Keywords in this Reference** | |
| <https://ieeexplore>.ieee.org/document/8070805 | | Mohan Ghai  Shamit Lal  Shivam Duggal  Shrey Manik | | | | Berlin database, Emotion recognition, Gradient boosting, MFCC, SVM, Random Forest | |
| **The Name of the Current Solution (Technique/ Method/ Scheme/ Algorithm/ Model/ Tool/ Framework/ ... etc )** | | **The Goal (Objective) of this Solution & What is the problem that need to be solved** | | | | **What are the components of it?** | |
| Random Forest | | The goal of the paper is to recognize emotions in speech signals and classify them into seven emotion output classes using machine learning techniques. | | | | Author used Mel Frequency Cepstral Coefficients (MFCC) and Berlin database of emotional speech. The components of the paper also included feature vector, classifiers which are used to recognize emotions in speech signals. | |
| **The Process (Mechanism) of this Work; Means How the Problem has Solved & Advantage & Disadvantage of Each Step in This Process** | | | | | | | |
| |  |  |  |  | | --- | --- | --- | --- | |  | **Process Steps** | **Advantage** | **Disadvantage (Limitation)** | | **1** | Feature Extraction | High accuracy in recognizing emotions in speech signals. | Only limited to 7 human emotions. | | **2** | Three Classification Algorithms are used for classifying an audio signal into one of the 7 classes. | Effective use of Mel Frequency Cepstral coefficients and energy of speech signals as feature inputs. | The classification algorithms wrongly predicted some of the samples belonging to happiness class as belonging to anger class. | | **3** | Results for each algorithm is summarized. | Potential for diverse applications in the field of interaction between humans and computers. | Increased computational complexity and risk of overfitting. | | | | | | | | |
| **Major Impact Factors in this Work** | | | | | | | |
| |  |  |  |  | | --- | --- | --- | --- | | **Dependent Variable** | **Independent Variable** | **Moderating variable** | **Mediating (Intervening ) variable** | | Emotion classification accuracy | MFCC, Energy of speech signals | Characteristics of the Berlin database of emotional speech | The process through which Mel Frequency Cepstral coefficients (MFCC) and energy of speech signals influence the relationship between the classification algorithms | | | | | | | | |
| |  | | --- | | **Relationship Among The Above 4 Variables in This article** | | In emotion classification accuracy, MFCC and energy of speech signals interact with the characteristics of the Berlin database, shaping their impact. Additionally, MFCC and energy mediate the influence of classification algorithms on accuracy. | | | | | | | | |
| **Input and Output** | | | **Feature of This Solution** | | | | **Contribution in This Work** |
| |  |  | | --- | --- | | **Input** | **Output** | | Emotional speech | Detected Emotion | | | | The system leverages perceptual features including Mel Frequency Cepstral coefficients (MFCC), energy of the speech signals, and the classification algorithms Support Vector Machine (SVM), Random Decision Forest, and Gradient Boosting. | | | | Considering the different classification strategies, the maximum accuracy is obtained for the database by using Random Decision Forest classifier. |
| **Positive Impact of this Solution in This Project Domain** | | | | | **Negative Impact of this Solution in This Project Domain** | | |
| Random Forest and Ensemble methods like gradient boosting are good advancement to improve performance and accuracy. | | | | | Speech emotion recognition models are always not accurate. They can be fooled by accents, background noise, and other factors. This could lead to misunderstandings and misinterpretations of people’s emotions. | | |
| **Analyse This Work By Critical Thinking** | | | | **The Tools That Assessed this Work** | | | **What is the Structure of this Paper** |
| Logically this is a good step that detects emotions in a speech and tests accuracy. | | | | Support Vector Machine (SVM), Random Forest, Gradient boosting. | | | Abstract   1. Introduction 2. Berlin Database of Emotional Speech 3. Related Work 4. Speech Emotion Recognition Framework 5. Experiment Results 6. Conclusion 7. Future Work |
| **Diagram/Flowchart** | | | | | | | |
|  | | | | | | | |

|  |
| --- |
| **13** |
| **Reference in APA format** | |  | | | | | |
| **URL of the Reference** | | **Authors Names and Emails** | | | | **Keywords in this Reference** | |
| <https://www>.mdpi.com/1424-8220/22/6/2378 | | Apeksha Aggarwal  Akshat Srivastava  Ajay Agarwal  Nidhi Chahal  Dilbag Singh  Abeer Ali Alnuaim  Aseel Alhadlaq  Heung-No Lee | | | | Speech emotion recognition, machine learning, Principal Component Analysis (PCA), deep neural network (DNN), feature extraction. | |
| **The Name of the Current Solution (Technique/ Method/ Scheme/ Algorithm/ Model/ Tool/ Framework/ ... etc )** | | **The Goal (Objective) of this Solution & What is the problem that need to be solved** | | | | **What are the components of it?** | |
| Two-way feature extraction | | The goal of the system is to improve speech emotion recognition by exploring two different methods of feature extraction. | | | | The system comprises two-way feature extraction methods, Principal Component Analysis (PCA), Deep Neural Network (DNN) and pre-trained VGG-16 model, multimodal speech data. | |
| **The Process (Mechanism) of this Work; Means How the Problem has Solved & Advantage & Disadvantage of Each Step in This Process** | | | | | | | |
| |  |  |  |  | | --- | --- | --- | --- | |  | **Process Steps** | **Advantage** | **Disadvantage (Limitation)** | | **1** | Two datasets are used Toronto Emotional Speech Set (TESS), Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS). | The extracted features are processed using principal component analysis (PCA) and a deep neural network (DNN) with dense and dropout layers. This combination of feature extraction and modelling techniques helps improve the accuracy of speech emotion recognition. | Increase in complexity of the system | | **2** | Two approaches are introduced for extracting features: i) working directly on the audio dataset to obtain numerical features, ii) utilized spectrograms as image features. | By utilizing super convergence, the method extracts two sets of potential features from the speech data. This approach allows for a more comprehensive representation of the emotional content in the speech signals. | Dependency on pre-trained models | | **3** | The VGG-16 model outputs a feature vector. The feature vector is classified into one of the seven emotions. | It also involves multimodal data integration. | Sensitive to input quality and limited interpretability. | | | | | | | | |
| **Major Impact Factors in this Work** | | | | | | | |
| |  |  |  |  | | --- | --- | --- | --- | | **Dependent Variable** | **Independent Variable** | **Moderating variable** | **Mediating (Intervening ) variable** | | Speech emotion recognition accuracy | Two-way feature extraction methods | Dataset used (RAVDESS vs numeric features on a DNN) | Process through which two-way feature extraction methods super convergence. | | | | | | | | |
| |  | | --- | | **Relationship Among The Above 4 Variables in This article** | | In speech emotion recognition accuracy, two-way feature extraction methods interact with the dataset used, either RAVDESS or numeric features on a DNN, which moderates their impact. Additionally, the process of super convergence through these methods mediates their influence on accuracy. | | | | | | | | |
| **Input and Output** | | | **Feature of This Solution** | | | | **Contribution & The Value of This Work** |
| |  |  | | --- | --- | | **Input** | **Output** | | Raw audio signal | Finding the emotion of the speech | | | | Feature extraction plays a crucial role in the process of recognizing speech emotions. Utilizing PCA for feature extraction with DNN and utilizing pre-trained VGG-16 model for speech emotion recognition. | | | | To the extent this work is designed for the two -way feature extraction method for Speech Emotion Recognition. |
| **Positive Impact of this Solution in This Project Domain** | | | | | **Negative Impact of this Solution in This Project Domain** | | |
| The proposed SER method improves feature extraction for emotional communication and excels in accuracy, adaptability, and practicality. | | | | | Negative impact of this solution include dependency on dataset characteristics, regional bias, and varied accuracy across emotions. | | |
| **Analyse This Work By Critical Thinking** | | | | **The Tools That Assessed this Work** | | | **What is the Structure of this Paper** |
| Since this is designed involving Deep learning and feature extraction from both audio and Mel spectrograms which made a valuable addition to the field, the method’s evaluation on a single benchmark dataset raises concerns about its generalizability. | | | | Accuracy, f1 score. | | | Abstract   1. Introduction 2. Materials and Methods 3. Results 4. Conclusion |
| **Diagram/Flowchart** | | | | | | | |
|  | | | | | | | |

|  |
| --- |
| **14** |
| **Reference in APA format** | |  | | | | | |
| **URL of the Reference** | | **Authors Names and Emails** | | | | **Keywords in this Reference** | |
| <https://ieeexplore>.ieee.org/document/9528931 | | Tedd Kourkounakis  Amirhossein Hajavi  Ali Etemad | | | | Stuttering, deep learning, disfluency, Mel-frequency spectral coefficients (MFCC), Bidirectional Long short term memory(BLSTM), Squeeze-and-Excitation residual (SE-ResNet). | |
| **The Name of the Current Solution (Technique/ Method/ Scheme/ Algorithm/ Model/ Tool/ Framework/ ... etc )** | | **The Goal (Objective) of this Solution & What is the problem that need to be solved** | | | | **What are the components of it?** | |
| FluentNet | | To detect and classify different forms of stuttering and speech disfluencies. | | | | The system components encompass Squeeze-and-Excitation Residual Convolutional Neural Network, Bidirectional Long Short-Term Memory (LSTM) Layers and Attention mechanism | |
| **The Process (Mechanism) of this Work; Means How the Problem has Solved & Advantage & Disadvantage of Each Step in This Process** | | | | | | | |
| |  |  |  |  | | --- | --- | --- | --- | |  | **Process Steps** | **Advantage** | **Disadvantage (Limitation)** | | **1** | Collecting speech samples from UCLASS and LibriStutter dataset. | Provides End-to-end solution | Limited dataset size, computationally expensive. | | **2** | Generating spectrogram from audio clips. | Spectral Frame-level representations | Challenges in classifying interjections | | **3** | Utilizing a Squeeze-and-Excitation Residual Network (SE-ResNet) and passing spectrogram feature vectors through bidirectional LSTM layers. | Incorporates temporal relationships | Developers of FluentNet do not have a clear privacy policy in place, which raises further concerns about how user data is being used. | | **4** | Training model on annotated dataset and evaluate performance. Analyzing experimental results for accuracy. | Achieves state-of-the-art performance. |  | | | | | | | | |
| **Major Impact Factors in this Work** | | | | | | | |
| |  |  |  |  | | --- | --- | --- | --- | | **Dependent Variable** | **Independent Variable** | **Moderating variable** | **Mediating (Intervening ) variable** | | Stutter detection and recognition performance | FluentNet architecture, stutter types | Characteristics of datasets | Learning of strong spectral frame-level representations | | | | | | | | |
| |  | | --- | | **Relationship Among The Above 4 Variables in This article** | | In stutter detection and recognition performance, the FluentNet architecture and types of stutter serve as the primary focus. The characteristics of datasets moderate their impact, while the learning of strong spectral frame-level representations acts as a mediating factor, influencing how the FluentNet architecture and stutter types affect performance. | | | | | | | | |
| **Input and Output** | | | **Feature of This Solution** | | | | **Contribution & The Value of This Work** |
| |  |  | | --- | --- | | **Input** | **Output** | | Short-Time Fourier Transform (STFT) spectrograms of audio clips | Types of stutters (sound repetition, word repetition, prolongation etc) | | | | Features of the system include Squeeze-and-Excitation Residual Network, Bidirectional Long Short-Term Memory (BLSTM) Layers and global attention mechanism for stuttering detection. | | | | This model introduces a novel method of stutter detection and it addresses potential errors and computational complexity associated with ASR. |
| **Positive Impact of this Solution in This Project Domain** | | | | | **Negative Impact of this Solution in This Project Domain** | | |
| The authors found that using this model could potentially be used to develop smart and interactive tools for detection and therapy, as well as to improve presentation skills. | | | | | The FluentNet model could misdiagnose people with stuttering, which could lead to them receiving inappropriate treatment. | | |
| **Analyse This Work By Critical Thinking** | | | | **The Tools That Assessed this Work** | | | **What is the Structure of this Paper** |
| This work innovatively addresses stutter detection by directly analysing audio signals, eliminating reliance on automatic speech recognition. The integration of Squeeze-and-Excitation Residual Network (SE-ResNet) and bidirectional LSTM layers contributes to effective feature learning, providing potential advancements in capturing variety speech patterns. | | | | FluentNet, a deep neural network, and several baseline and state-of-the-art techniques. | | | Abstract   1. Introduction 2. Related Work 3. Proposed Method 4. Experiments 5. Conclusion |
| **Diagram/Flowchart** | | | | | | | |
|  | | | | | | | |

|  |
| --- |
| **15** |
| **Reference in APA format** | |  | | | | | |
| **URL of the Reference** | | **Authors Names and Emails** | | | | **Keywords in this Reference** | |
| <https://arxiv>.org/abs/2002.07590 | | Manas Jain  Shruthi Narayan  Pratibha Balaji  Bharath K P  Abhijit Bhowmick  Karthik R  Rajesh Kumar Muthu | | | | Emotion, Support Vector Machine (SVM), Mel Frequency Ceptral Coefficients (MFCC), feature extraction, LPCC. | |
| **The Name of the Current Solution (Technique/ Method/ Scheme/ Algorithm/ Model/ Tool/ Framework/ ... etc )** | | **The Goal (Objective) of this Solution & What is the problem that need to be solved** | | | | **What are the components of it?** | |
| Support Vector Machine | | The goal of this work is to address the identification and classification of speech into various emotions. | | | | The systems components include the input speech signal, feature extraction using MFCC and LPC, classification based on SVM, and the output. | |
| **The Process (Mechanism) of this Work; Means How the Problem has Solved & Advantage & Disadvantage of Each Step in This Process** | | | | | | | |
| |  |  |  |  | | --- | --- | --- | --- | |  | **Process Steps** | **Advantage** | **Disadvantage (Limitation)** | | **1** | The author used two datasets, LDC and UGA as input. | SVM-based SER system has demonstrated high accuracy in recognizing emotions from speech utterances. | The system’s training had very less amount of labelled speech data. | | **2** | Extracted acoustic and prosodic features from speech recordings and trained an SVM classifier using the selected features and emotion labels. | This system is robust to noise and variations in recording conditions, and can be adapted to different languages and accents. | The model is Computationally expensive. | | **3** | Evaluated the trained SVM classifier on a separate test dataset to assess its performance. |  |  | | | | | | | | |
| **Major Impact Factors in this Work** | | | | | | | |
| |  |  |  |  | | --- | --- | --- | --- | | **Dependent Variable** | **Independent Variable** | **Moderating variable** | **Mediating (Intervening ) variable** | | Emotion classification | Acoustic features and prosodic features | Speaker characteristics | Contextual information | | | | | | | | |
| |  | | --- | | **Relationship Among The Above 4 Variables in This article** | | For emotion classification, acoustic and prosodic features play a central role, while speaker characteristics moderate their impact. Contextual information further influences how these features contribute to the classification process. | | | | | | | | |
| **Input and Output** | | | **Feature of This Solution** | | | | **Contribution & The Value of This Work** |
| |  |  | | --- | --- | | **Input** | **Output** | | Raw audio recording of speech | Emotion Classification | | | | The SVM-based SER system demonstrates high accuracy and robustness in recognizing emotions from speech utterances. Its ability to handle non-linear relationships and adapt to different languages further enhances its versatility. | | | | The SVM-based SER system offers a novel approach to speech emotion recognition, combining support vector machines with acoustic and prosodic feature extraction to achieve high accuracy and robustness in real-world applications. |
| **Positive Impact of this Solution in This Project Domain** | | | | | **Negative Impact of this Solution in This Project Domain** | | |
| This model has improved customer satisfaction by accurately recognizing customer emotions, tracks changes in emotional patterns over time. | | | | | Misinterpretation of emotions could lead to inappropriate responses, misunderstandings, and further emotional distress. | | |
| **Analyse This Work By Critical Thinking** | | | | **The Tools That Assessed this Work** | | | **What is the Structure of this Paper** |
| The SVM-based SER system demonstrates promising advancements in speech emotion recognition, exhibiting high accuracy, robustness to noise, and the ability to handle non-linear relationships between features and emotions. However, it is essential to critically evaluate its potential limitations and broader implications. | | | | Confusion matrix, f1-score | | | Abstract   1. Introduction 2. Methodology 3. SVM Algorithm 4. Datasets 5. Simulation outputs and results 6. Conclusion and Future Works |
| **Diagram/Flowchart** | | | | | | | |
|  | | | | | | | |

|  |
| --- |
| **16** |
| **Reference in APA format** | |  | | | | | |
| **URL of the Reference** | | **Authors Names and Emails** | | | | **Keywords in this Reference** | |
| <https://arxiv>.org/abs/2105.05599 | | md.sahidullah | | | | Speech Recognition, Self-Supervised Learning ,Speech Representations, Automatic Speech Recognition (ASR), Language Modeling | |
| **The Name of the Current Solution (Technique/ Method/ Scheme/ Algorithm/ Model/ Tool/ Framework/ ... etc )** | | **The Goal (Objective) of this Solution & What is the problem that need to be solved** | | | | **What are the components of it?** | |
| StutterNet: Stuttering Detection Using Time Delay Neural Network | | The aim of the paper is to propose a self-supervised learning approach for training speech representations. Speech representations are features that can be used to represent speech signals. | | | | Data preparation, Self-supervised learning objective, ASR training procedure | |
| **The Process (Mechanism) of this Work; Means How the Problem has Solved & Advantage & Disadvantage of Each Step in This Process** | | | | | | | |
| |  |  |  |  | | --- | --- | --- | --- | |  | **Process Steps** | **Advantage** | **Disadvantage (Limitation)** | | **1** | The authors use a variety of techniques to prepare the data for self-supervised learning, including data augmentation, noise injection, and masking. | The proposed approach has been shown to improve the performance of ASR systems, even when the ASR systems are trained on a small amount of labeled data. | The proposed approach is sensitive to the choice of hyperparameters, such as the learning rate and the batch size. | | **2** | The authors propose a self-supervised learning objective that is based on contrastive learning. In contrastive learning, the model learns to distinguish between positive and negative pairs of data. | The proposed approach reduces the amount of labeled data required to train ASR systems. This is important because labeled data is expensive and time-consuming to collect. | The proposed approach is prone to overfitting, especially when trained on a small amount of data. | | | | | | | | |
| **Major Impact Factors in this Work** | | | | | | | |
| |  |  |  |  | | --- | --- | --- | --- | | **Dependent Variable** | **Independent Variable** | **Moderating variable** | **Mediating (Intervening ) variable** | | The primary focus of the research is on the accuracy of stuttering detection, which serves as the dependent variable. | StutterNet System Implementation including the use of TDNN(Time Delay Neural Network), acoustic features like MFCCs, and the specific architecture design. | Factors influencing the relationship between the independent variable and the dependent variable | The optimization variables act as potential mediating variables, influencing the relationship between the StutterNet system implementation (independent variable) and stuttering detection accuracy (dependent variable). | | | | | | | | |
| |  | | --- | | **Relationship Among The Above 4 Variables in This article** |   In this research, the StutterNet system’s implementation (independent variable) directly influences the accuracy of stuttering detection (dependent variable). Optimization variables, such as layer size, context window, and filter bank size, serve as potential mediators, affecting the relationship between system implementation and detection accuracy. | | | | | | | |
| **Input and Output** | | | **Feature of This Solution** | | | | **Contribution & The Value of This Work** |
| |  |  | | --- | --- | | **Input** | **Output** | | The input to the solution is unlabeled speech data. This data can be in any language and can be noisy. | The output of the solution is a set of speech representations. | | | | The key feature of the solution is that it uses self-supervised learning to train speech representations. This means that the model learns from unlabeled data, without the need for human-transcribed transcripts. | | | | The main contribution of the paper is the proposal of a new self-supervised learning objective for ASR. This objective is based on contrastive learning, where the model learns to distinguish between positive and negative pairs of data. |
| **Positive Impact of this Solution in This Project Domain** | | | | | **Negative Impact of this Solution in This Project Domain** | | |
| The proposed solution has been shown to improve the performance of ASR systems, even when the ASR systems are trained on a small amount of labeled data. | | | | | Since this is a performance evaluation of various algorithms, not much to project on negative side as all the things used are defined in advance. | | |
| **Analyse This Work By Critical Thinking** | | | | **The Tools That Assessed this Work** | | | **What is the Structure of this Paper** |
| The research paper “Self-Supervised Speech Recognition with Speech Representations from ASR” proposes a new self-supervised learning objective for ASR. | | | | Self-supervised learning | | | Abstract   1. INTRODUCTION 2. RELATED WORK 3. PROPOSED ARCHITECTURE 4. EXPERIMENTAL EVALUATION 5. RESULTS 6. CONCLUSION |
| **Diagram/Flowchart** | | | | | | | |
|  | | | | | | | |

|  |
| --- |
| **17** |
| **Reference in APA format** | |  | | | | | |
| **URL of the Reference** | | **Authors Names and Emails** | | | | **Keywords in this Reference** | |
| <https://arxiv>.org/abs/2302.09044 | | Colin Lea  colin\_lea@apple.com | | | | speech input, accessibility, stuttering, voice assistants, dictation | |
| **The Name of the Current Solution (Technique/ Method/ Scheme/ Algorithm/ Model/ Tool/ Framework/ ... etc )** | | **The Goal (Objective) of this Solution & What is the problem that need to be solved** | | | | **What are the components of it?** | |
| Hybrid Decision Tree and Logistic Regression | | The Goal (Objective) demonstrate how many common errors can be prevented, resulting in a system that cuts utterances of 79.1% less often and improves word error rate from 25.4% to 9.9%. | | | | Propose technical solutions to improve the performance of speech recognition systems for people who stutter. | |
| **The Process (Mechanism) of this Work; Means How the Problem has Solved & Advantage & Disadvantage of Each Step in This Process** | | | | | | | |
| |  |  |  |  | | --- | --- | --- | --- | |  | **Process Steps** | **Advantage** | **Disadvantage (Limitation)** | | **1** | The authors conducted two surveys to understand the needs and experiences of people who stutter when using speech recognition systems. | The authors' proposed solutions resulted in a 79.1% reduction in the number of times utterances were cut off and a 15.5% improvement in word error rate. | The authors' work is still in the early stages of development, and more research is needed to validate their findings and develop robust and scalable solutions. | | **2** | They also conducted technical investigations to identify and address common errors in speech recognition systems. | The authors’ work could significantly improve the performance of speech recognition systems for people who stutter, making it easier for them to communicate and participate in everyday activities. | The authors' work Is also limited to a specific type of speech recognition system, and it is not clear how well their solutions would generalize to other types of systems. | | **3** | The authors then proposed three technical solutions to improve the performance of speech recognition systems for people who stutter. |  |  | | **4** | The authors evaluated their proposed solutions on a real-world dataset of speech from people who stutter. |  |  | | | | | | | | |
| **Major Impact Factors in this Work** | | | | | | | |
| |  |  |  |  | | --- | --- | --- | --- | | **Dependent Variable** | **Independent Variable** | **Moderating variable** | **Mediating (Intervening ) variable** | | The dependent variable is the Word Error Rate (WER), which is the key metric used to measure the accuracy of speech recognition. | Independent variables is the threshold set for endpoint detection in speech recognition. | Moderating variable is Stuttering severity may moderate the impact of interventions. | A mediating variable is Annotations highlighting dysfluency types in the speech. | | | | | | | | |
| |  | | --- | | **Relationship Among The Above 4 Variables in This article** | | In the context of speech recognition for people who stutter, the study investigates the impact of independent variables (Endpointer Threshold, ASR Decoder Tuning, Dysfluency Refinement) on dependent variables (Word Error Rate, Intent Error Rate), with stuttering severity serving as a moderating factor and dysfluency annotations as a mediating variable. | | | | | | | | |
| **Input and Output** | | | **Feature of This Solution** | | | | **Contribution in This Work** |
| |  |  | | --- | --- | | **Input** | **Output** | | The input to the system is a speech signal from a person who stutters. | The output of the system is a transcript of the speech signal. | | | | The key feature of the system is that it is robust to dysfluencies, such as repetitions and prolongations. | | | | The system can help people who stutter to communicate more effectively using speech recognition systems. |
| **Positive Impact of this Solution in This Project Domain** | | | | | **Negative Impact of this Solution in This Project Domain** | | |
| The proposed solution can make speech recognition systems more accessible to people who stutter. This is because the solution is robust to dysfluencies, such as repetitions and prolongations, which are common in stuttering speech. | | | | | The proposed solution may reduce the accuracy of speech recognition systems in non-stuttering speech. This is because the solution is tuned to detect and correct dysfluencies, which are not present in non-stuttering speech. | | |
| **Analyse This Work By Critical Thinking** | | | | **The Tools That Assessed this Work** | | | **What is the Structure of this Paper** |
| It involves assessing the innovative approaches used to enhance speech recognition for people who stutter. | | | | The tools employed for evaluation include dysfluency annotations, ASR models with varied architectures, and statistical methods such as Wilcoxon signed rank tests. | | | Abstract   1. Introduction 2. Related Work 3. Proposed Method 4. Experiment Results 5. Conclusion |
| **Diagram/Flowchart** | | | | | | | |
|  | | | | | | | |

|  |
| --- |
| **18** |
| **Reference in APA format** | |  | | | | | |
| **URL of the Reference** | | **Authors Names and Emails** | | | | **Keywords in this Reference** | |
| <https://arxiv>.org/abs/2106.06598 | | Shinji Watanabe | | | | speech sentiment analysis, pre-trained language model, pseudo label-based semi-supervised training speech sentiment analysis, pre-trained language model, pseudo label-based semi-supervised training | |
| **The Name of the Current Solution (Technique/ Method/ Scheme/ Algorithm/ Model/ Tool/ Framework/ ... etc )** | | **The Goal (Objective) of this Solution & What is the problem that need to be solved** | | | | **What are the components of it?** | |
| A Proposed Model for Leveraging Pre-trained Language Model for Speech Sentiment Analysis | | The objective of the research paper is to propose a novel approach for speech sentiment analysis using pre-trained language models and a pseudo label-based semi-supervised training strategy. An approach to transfer knowledge from the written text to spoken text or speech domain using an LM | | | | MFCCs are a type of audio feature that is commonly used in speech processing. MFCCs are extracted from the speech signal and fed into the pre-trained language model. | |
| **The Process (Mechanism) of this Work; Means How the Problem has Solved & Advantage & Disadvantage of Each Step in This Process** | | | | | | | |
| |  |  |  |  | | --- | --- | --- | --- | |  | **Process Steps** | **Advantage** | **Disadvantage (Limitation)** | | **1** | 2-step pipeline approach employs Automatic Speech Recognition (ASR) and transcripts-based sentiment analysis separately. | They can be used to train a sentiment analysis system on a large amount of data without human sentiment annotation. | They require a pre-trained language model, which can be expensive to train and deploy. | | **2** | Pseudo label-based semi-supervised training strategy uses a language model on an end-to-end speech sentiment approach. | They can be used to train a sentiment analysis system that is robust to noise in the speech signal. | The proposed approaches may not be as effective for speech sentiment analysis in low-resource languages or for speech sentiment analysis of specific domains, such as medical or legal speech. | | | | | | | | |
| **Major Impact Factors in this Work** | | | | | | | |
| |  |  |  |  | | --- | --- | --- | --- | | **Dependent Variable** | **Independent Variable** | **Moderating variable** | **Mediating (Intervening ) variable** | | recall, precision, and F1 scores. | Text Source: ASR transcripts as input data | Quantity of Labeled Data: Labeled data acts as a moderating variable in the relationship between the independent variables and model performance. | Pseudo Label-based Semi-Supervised Training: Pseudo label-based semi-supervised training as a mediating variable between theeature variables and the dependent variable. | | | | | | | | |
| |  | | --- | | **Relationship Among The Above 4 Variables in This article** | | In contemporary sentiment analysis research, the literature underscores the pivotal role of pre-trained language models, notably BERT, as an independent variable. These models significantly enhance the performance of speech sentiment analysis, representing a key focus in the exploration of cutting-edge methodologies for efficient sentiment classification in spoken language datasets. | | | | | | | | |
| **Input and Output** | | | **Feature of This Solution** | | | | **Contribution & The Value of This Work** |
| |  |  | | --- | --- | | **Input** | **Output** | | Speech signal | Sentiment | | | | The system uses a pseudo label-based semi-supervised training strategy to reduce the need for human-labeled data. | | | | The system can be used to develop customer service chatbots that can understand and respond to customer queries in a more effective manner. |
| **Positive Impact of this Solution in This Project Domain** | | | | | **Negative Impact of this Solution in This Project Domain** | | |
| The proposed solution can be used to improve the performance of speech-based applications, such as customer service chatbots and virtual assistants. This is because the system can extract features from the speech signal that are informative for sentiment analysis. | | | | | The proposed solution may reduce the accuracy of speech-based applications in certain cases. For example, the system may not be as effective for speech sentiment analysis in low-resource languages or for speech sentiment analysis of specific domains, such as medical or legal speech. | | |
| **Analyse This Work By Critical Thinking** | | | | **The Tools That Assessed this Work** | | | **What is the Structure of this Paper** |
| The research paper is well-written and the proposed approaches are well-motivated. The authors evaluate their approaches on a real-world dataset and show that they can improve F1 scores consistently compared to systems without a language model. The authors also acknowledge the limitations of their work and suggest directions for future research. | | | | The paper used these tools to evaluate their proposed approaches on a real-world dataset of speech. The results showed that the proposed approaches improved F1 scores consistently compared to systems without a language model. The confusion matrices showed that the proposed approaches were able to reduce the number of misclassified instances. | | | Abstract   1. Introduction 2. Related work 3. Approaches 4. Experiments 5. Conclusion |
| **Diagram/Flowchart** | | | | | | | |
|  | | | | | | | |

|  |
| --- |
| **19** |
| **Reference in APA format** | |  | | | | | |
| **URL of the Reference** | | **Authors Names and Emails** | | | | **Keywords in this Reference** | |
| <https://www>.mdpi.com/1424-8220/22/17/6369 | | Bagus Tris Atmaja and Akira Sasou | | | | affective computing; sentiment analysis; speech emotion recognition; sentiment analysis and emotion recognition; universal speech representation | |
| **The Name of the Current Solution (Technique/ Method/ Scheme/ Algorithm/ Model/ Tool/ Framework/ ... etc )** | | **The Goal (Objective) of this Solution & What is the problem that need to be solved** | | | | **What are the components of it?** | |
| Sentiment Analysis and Emotion Recognition from Speech Using Universal Speech Representations | | The objective of the research paper is to conduct a comprehensive review of the state-of-the-art in sentiment analysis and emotion recognition from speech using universal speech representations (USRs). | | | | It covers a wide range of topics, including different approaches to USR extraction, different methods for applying USRs to sentiment analysis and emotion recognition, and the challenges and limitations of the current state-of-the-art. | |
| **The Process (Mechanism) of this Work; Means How the Problem has Solved & Advantage & Disadvantage of Each Step in This Process** | | | | | | | |
| |  |  |  |  | | --- | --- | --- | --- | |  | **Process Steps** | **Advantage** | **Disadvantage (Limitation)** | | **1** | The first step is to extract USRs from the speech signal. This can be done using a variety of approaches, such as deep learning-based approaches and traditional signal processing-based approaches. | USRs are designed to be universal, meaning that they can be used for a variety of speech processing tasks, including sentiment analysis and emotion recognition. This makes them a versatile tool for developing speech-based applications. | There is a lack of large-scale datasets that are labeled for both speech and sentiment or emotion. This can make it difficult to train and evaluate sentiment analysis and emotion recognition systems that use USRs. | | **2** | Once the USRs have been extracted, they can be applied to sentiment analysis or emotion recognition using a variety of methods, such as machine learning methods and deep learning methods. | USRs are typically more robust to noise and other variations in the speech signal than traditional speech features. This makes them more suitable for real-world applications. | USRs may not be as effective for sentiment analysis and emotion recognition in certain domains, such as medical speech or legal speech. This is because the speech characteristics in these domains may be different from those in general-purpose speech datasets. | | | | | | | | |
| **Major Impact Factors in this Work** | | | | | | | |
| |  |  |  |  | | --- | --- | --- | --- | | **Dependent Variable** | **Independent Variable** | **Moderating variable** | **Mediating (Intervening ) variable** | | For each sentiment analysis and emotion recognition task, the accuracy scores can be considered dependent variables. | Independent variable could be the number of classes in sentiment analysis tasks | A moderating variable could be the size of the datasets used for training the models. | A mediating variable could be the acoustic features extracted by the model, which serve as an intermediary step in the process of sentiment analysis and emotion recognition from speech. | | | | | | | | |
| |  | | --- | | **Relationship Among The Above 4 Variables in This article** | | In the research paper, sentiment analysis and emotion recognition tasks are evaluated (dependent variables) using different UniSpeech-SAT models (independent variables). The dataset size may moderate performance, while acoustic features extracted by the models (mediating variables) contribute to understanding the complex relationship between model architecture and task outcomes. | | | | | | | | |
| **Input and Output** | | | **Feature of This Solution** | | | | **Contribution & The Value of This Work** |
| |  |  | | --- | --- | | **Input** | **Output** | | The input to the solution is a speech signal. The speech signal can be in any language and can be noisy. | The output of the solution is the sentiment or emotion of the speech signal. The sentiment can be positive, negative, or neutral. The emotion can be happiness, sadness, anger, fear, or surprise. | | | | The key feature of the solution is the use of universal speech representations (USRs). USRs are a type of speech representation that is designed to be universal, meaning that they can be used for a variety of speech processing tasks, including sentiment analysis and emotion recognition. USRs are typically more robust to noise and other variations in the speech signal than traditional speech features, and they have been shown to achieve state-of-the-art results on a variety of sentiment analysis and emotion recognition benchmarks. | | | | The solution has the potential to improve significantly the performance of speech-based applications, such as customer service chatbots, virtual assistants, and medical diagnostic systems. By using USRs, these applications can better understand the sentiment and emotion of the user, which can lead to more accurate and effective responses. |
| **Positive Impact of this Solution in This Project Domain** | | | | | **Negative Impact of this Solution in This Project Domain** | | |
| The solution can be used to develop customer service chatbots that can better understand the sentiment and emotion of the customer, which can lead to more accurate and effective responses. This can improve customer satisfaction and reduce the cost of customer service. | | | | | Theeaturion can be used to develop virtual assistants that can better understand the user’s intent and emotion, which can lead to more personalized and engaging experiences. This can increase user engagement and adoption of virtual assistants. | | |
| **Analyse This Work By Critical Thinking** | | | | **The Tools That Assessed this Work** | | | **What is the Structure of this Paper** |
| The paper is well-written and informative. The paper provide a good overview of the state-of-the-art in USRs, sentiment analysis, and emotion recognition. They also discuss the potential of USRs to improve significantly the performance of sentiment analysis and emotion recognition systems. However, the paper does not provide a detailed evaluation of the proposed solution, nor does it discuss the potential negative impacts of the solution in detail. | | | | The research paper on sentiment analysis and emotion recognition from speech using universal speech representations (USRs) does not provide a detailed evaluation of the proposed solution. However, the authors mention that they are planning to evaluate the solution on a variety of datasets in future work. | | | Abstract   1. Introduction 2. Related Work 3. Methods 4. Experiments 5. Results and Discussion |
| **Diagram/Flowchart** | | | | | | | |
|  | | | | | | | |

|  |
| --- |
| **20** |
| **Reference in APA format** | |  | | | | | |
| **URL of the Reference** | | **Authors Names and Emails** | | | | **Keywords in this Reference** | |
| <https://www>.ijert.org/stuttered-speech-recognition-using-convolutional-neural-networks | | Phani Bhushan S | | | | Stuttered Speech Recognition (SSR), Convolution Neural Network (CNN), Mel Frequency Co-efficient (MFCC). | |
| **The Name of the Current Solution (Technique/ Method/ Scheme/ Algorithm/ Model/ Tool/ Framework/ ... etc )** | | **The Goal (Objective) of this Solution & What is the problem that need to be solved** | | | | **What are the components of it?** | |
| Stuttered Speech Recognition using Convolutional Neural Networks | | The goal of the research paper is to develop a system that can recognize stuttered speech using convolutional neural networks (CNNs). CNNs are a type of machine learning algorithm that is well-suited for tasks such as image recognition and speech recognition. | | | | The paper use a variety of features to represent the stuttered speech signals, including Mel-frequency cepstral coefficients (MFCCs), pitch, and energy. | |
| **The Process (Mechanism) of this Work; Means How the Problem has Solved & Advantage & Disadvantage of Each Step in This Process** | | | | | | | |
| |  |  |  |  | | --- | --- | --- | --- | |  | **Process Steps** | **Advantage** | **Disadvantage (Limitation)** | | **1** | The stuttered speech data is prepared by performing data augmentation, noise injection, and masking. This helps to improve the robustness of the system to noise and other variations in the speech signal. | It achieves high accuracy in distinguishing between stuttered and non-stuttered speech. | It requires a large amount of data to train the CNN. | | **2** | A variety of features are extracted from the stuttered speech signals, including MFCCs, pitch, and energy. These features are used to represent the stuttered speech signals in a way that can be easily processed by the CNN. | It is robust to noise and other variations in the speech signal. | It is sensitive to the choice of hyperparameters, such as the learning rate and the batch size. | | **3** | The CNN is trained using the Adam optimizer and the cross-entropy loss function to distinguish between stuttered and non-stuttered speech. | It is computationally efficient and can be implemented in real time. | It is prone to overfitting, especially when trained on a small amount of data. | | **4** | The trained CNN is used to recognize stuttered speech in new data. |  |  | | | | | | | | |
| **Major Impact Factors in this Work** | | | | | | | |
| |  |  |  |  | | --- | --- | --- | --- | | **Dependent Variable** | **Independent Variable** | **Moderating variable** | **Mediating (Intervening ) variable** | | Measurement of accuracy in stuttered speech recognition serves as the dependent variable, reflecting the effectiveness of the proposed system. | The independent variable is the implementation of the system using a combination of Weighted Mel Frequency Cepstral Coefficient feature extraction and Convolutional Neural Networks. | Moderating variables such as the characteristics of the speech dataset, variations in speech patterns, or other contextual factors. | Identify any mediating variables that influence the relationship between the independent variable (SSR system) and the dependent variable (accuracy in stuttered speech recognition). | | | | | | | | |
| |  | | --- | | **Relationship Among The Above 4 Variables in This article** | | In this study, the independent variable is the implementation of a Stuttered Speech Recognition (SSR) system using CNN and Weighted MFCC. The dependent variable is the accuracy of stuttered speech recognition. Contextual factors, such as variations in speech datasets, act as moderating variables influencing system effectiveness. The underlying processes in the SSR system constitute mediating variables explaining the mechanism through which the system impacts accuracy. | | | | | | | | |
| **Input and Output** | | | **Feature of This Solution** | | | | **Contribution & The Value of This Work** |
| |  |  | | --- | --- | | **Input** | **Output** | | The input to the proposed stuttered speech recognition system is a stuttered speech signal. | The output of the system is a binary classification label, indicating whether the input speech signal is stuttered or non-stuttered. | | | | It uses a convolutional neural network (CNN) to extract features from the stuttered speech signal. CNNs are well-suited for this task because they are able to learn spatial and temporal patterns in the data. | | | | The system is robust to noise and other variations in the speech signal. This makes it suitable for use in real-world applications. |
| **Positive Impact of this Solution in This Project Domain** | | | | | **Negative Impact of this Solution in This Project Domain** | | |
| The system can be used to develop new speech recognition systems that are specifically designed for stuttering speakers. These systems could be used to improve the performance of speech recognition systems in noisy environments, to make it easier for stuttering speakers to control smart devices, and to develop new communication tools for people who stutter. | | | | | The system is prone to overfitting, especially when trained on a small amount of data. This means that the system may learn the characteristics of the training data too well and be unable to generalize to new data. | | |
| **Analyse This Work By Critical Thinking** | | | | **The Tools That Assessed this Work** | | | **What is the Structure of this Paper** |
| The research paper proposes a novel approach to stuttered speech recognition using convolutional neural networks (CNNs). The authors show that their proposed approach achieves high accuracy in distinguishing between stuttered and non-stuttered speech, even in noisy environments. | | | | Convolutional Neural Networks | | | 1. INTRODUCTION 2. LITERATURE SURVEY 3. PROPOSED METHODOLOGY 4. CONCLUSION |
| **Diagram/Flowchart** | | | | | | | |
|  | | | | | | | |

* 1. **COMPARISON TABLE**

|  |  |  |  |
| --- | --- | --- | --- |
| **Author** | **Year** | **Approach** | **Description** |
| [Tedd](https://ieeexplore.ieee.org/author/37088481517) [Kourkounakis](https://ieeexplore.ieee.org/author/37088481517), [Amirhossein](https://ieeexplore.ieee.org/author/37088482795) [Hajavi](https://ieeexplore.ieee.org/author/37088482795) and  [Ali Etemad](https://ieeexplore.ieee.org/author/37087028037) | 2020 | Deep residual  network with bidirectional long short-term memory layers. | A deep learning model using residual networks and bidirectional LSTMs accurately identifies different types of stutters in speech. This is a major advancement with potential for further improvements, like classifying even more types of disfluencies. |
| Elmar Noeth, Thomas Wittenberg and Michael Decher | 2000 | Hidden Markov Model (HMMs) | The solution proposed extracts features from the speech data and trained the HMMs to model the speech patterns of stuttered and non-stuttered speech. The trained models were then used to accurately recognize stuttering in new speech samples, as demonstrated by the high recognition rates obtained in the experiments. |
| S. Lalitha,  Shikha Tripathi and  Deepa Gupta | 2019 | Using Depp Neural Networks (DNN) in conjunction with diverse perceptual speech features for the task of speech emotion recognition (SER). | Deep neural networks  (FFBP) and Mel/Bark filter banks extract speaker-independent, compact features from audio, enabling  high-precision emotion detection. This combination offers a powerful tool for unlocking emotions hidden within audio signals. |
| [Bassam Ali](https://ieeexplore.ieee.org/author/38276950600) [Al-Qatab](https://ieeexplore.ieee.org/author/38276950600) and [Mumtaz Begum](https://ieeexplore.ieee.org/author/38241272200) [Mustafa](https://ieeexplore.ieee.org/author/38241272200) | 2021 | Automatic speech recognition (ASR) systems | This involves integrating four Acoustic features and employ seven feature selection methods to craft a hybrid approach. |
| Sadeen Alharbi. Madina Hasan,  Anthony J H Simons,  Shelagh Brumfitt and  Phil Green | 2017 | ML approaches, specifically the HELM and CRF | The solution’s key feature is its use of machine learning approaches (HELM and CRF) to detect and classify stuttering even since children speech transcripts, along with data agumentation. |
| Hadhami Aouani and Yassine Ben Ayed | 2020 | Feature-based SVM Emotion Recognition with Auto-Encoder-B ased Feature Dimension Reduction. | To improve the effectiveness of emotion recognition, a proposed approach will use an enhanced blend of features such as MFCC, ZCR, TEO and HNR. |
| I.Husbaan Attar, Nilesh K.Kadole, Omkar G, Karanjekar  Devang, R. Nagarkar | 2022 | ML approaches, Convolutional Neural Networks, Librosa library | A CNN trained on MFCC features extracted from audio (thanks to librosa) predicts emotions in new  samples, excelling in accuracy and speaker-independence  .Layered convolutional networks unlock  emotional nuances  hidden within audio, opening doors for advanced emotion recognition applications. |
| Dr. Mrs. Gresha Bhatia, Binoy  Saha, Mansi Khamkar and Ashish Chandwani , Reshma Khot | 2021 | Deep Learning, is employed specifically utilizing Mel Frequency Cepstral coefficients (MFCC),  Gated recurrent unit (GRU), Support Vector Machine (SVM) | The proposed solution is to develop a personalized stuttering therapy system that accurately diagnoses stutter and suggests appropriate training exercises for practice. |
| Seema Barda | 2019 | Automated Speech Recognition (ASR) employs Mel  Frequency (MFCC)for feature extraction, coupled with SVM for feature | The approach presented in the paper utilizes Support Vector Machine (SVM) as a pattern recognition method for the automatic assessment of stuttering severity in speech. |
| M. Mahendran, R. Visalakshi ,S. Balaji | 2021 | Convolutional Neural Networks | In the phonetically balanced articulatory phonology of healthy and dysthymic speakers, the model is used to assess frame level distance patterns. |
| Mustaqeem and Soonil Kwon | 2019 | Deep Stride Convolutional Neural Network (DSCNN) for speech emotion information. | According to this solution, enhanced speech is converted to spectrograms which then uses stride CNN architectures to extract features from the spectrograms after noise and silence signals are eliminated using the dynamic adaptive threshold methodology. |
| Mohan Ghai,  Shamit Lal,  Shivam Dugga and Shrey Manik | 2017 | Random Forest to distinguish between seven distinct emotions. | The Mel Frequency Cepstral Coefficients (MFCC) serve as the foundation for this solution. To train classification algorithms like  Support Vector Machine(SVM), Random Decision Forest, and Gradient Boosting to predict the proper emotions, the characteristics taken from speech are transformed into feature vectors. |
| Apeksha Aggarwal, Akshat Srivastava, Ajay Agarwal, Nidhi Chahal, Dilbag Singh, Abeer Ali Alnuaim,Aseel Alhadlaq and Heung-No Lee | 2022 | Two-way feature extraction and deep transfer learning. | This system investigated two distinct techniques for feature extraction. I) The first feature set is obtained by the application of Principal Component Analysis (PCA). After that, a deep neural network (DNN) featuring dropout and dense layers is put into practice. Ii) The pre-trained VGG-16 model receives the 2D pictures as Input for feature extraction after  Mel-Spectrogram features are retrieved from audio  recordings. |
| Tedd Kourkounakis, Amirhossein Hajavi, and Ali Etemad | 2021 | FluentNet for detecting several different disfluency types. | This solution consists of a set of bidirectional long short-term memory (BLSTM) layers that use an attention mechanism to focus on significant portions of speech that demonstrate the strong performance, followed by a squeeze-and-excitation residual convolutional neural  network that facilitates the learning of strong spectral frame level representations. |
| Manas Jain, Shruthi Narayan, Pratibha Balaji, Bharath KP, Abhijit Bhowmick, Karthik R, Rajesh Kumar Muthu | 2020 | Support Vector Machine (SVM), One Against All (OAA) and Gender Dependent Classification | The solution’s key feature is the use of SVM to gain higher classification accuracy. Gender dependent classifiers show better accuracy of 84.42%. Mel-Frequency Cepstral Coefficient has given higher accuracy for better feature extraction. |
| Suwon Shon,  Pablo Brusco, Jing Pan | 2021 | A two-stage pipeline approach implemented, involving Automatic Speech Recognition (ASR), sentiment analysis based on the transcripts conducted independently. | It introduces a two-step pipeline and an end-to-end approach, demonstrating improved F1 scores and reduced human supervision.  Experimental results highlight the robustness of BERT-based models to ASR errors and the potential for substantial savings in human annotation efforts. |
| Bagus Tris Atmaja , Akira Sasou | 2022 | A system built upon Time-Delay Neural Network (TDNN) | The research introduces StutterNet, a novel stuttering detection system based on a time-delay neural network. Unlike existing methods relying on automatic speech recognition, StutterNet utilizes acoustic signals, achieving promising results on the UCLASS stuttering dataset. |
| Shakeel, A.  Sheikh , Md Sahidullah , Fabrice Hirsch , Slim Ouni | 2021 | Detection using time-delay layers and  Mel-frequency cepstral coefficients | Proposed for stuttering detection using time-delay Layers and Mel-frequency cepstral coefficients. It outperforms existing methods, achieving a 4.69% gain in accuracy and 0.03 increase in MCC. |
| Colin Lea, Zifang Huang, Jaya Narain | 2022 | It identifies issues such as high word error rates and  truncation in existing models. | In this manuscript, we present survey findings gathered from 61 individuals who stuuter and conduct speech recognition experiments on a data set comprising 91 individuals.However irrespective of current usage over half of the respondents express a willingness to use speech technology more regularly if it demonstrated improved accuracy. |
| Phani Bhushan S, Vani H Y, D K Shivkumar | 2021 | In this approach, we have explored the scalability of Stuttered Speech recognition using Convolutional Neural Networks (CNN) | Stuttered Speech recognition (SSR) system is developed using CNN and weighted Mel Frequency Cepstral Coefficients. The research uses UCLASS dataset, achieving a recognition accuracy upto 92% in identifying stuttered speech. |

* 1. **WORK EVALUATION TABLE**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Work Goal** | **System’s Components** | **System’s Mechanism** | **Features /Characteristics** | **Performance** | **Advantages** | **Limitations /Disadvantages** | **Results** |
| **Hadhami Aouani**  **Yassine Ben Ayed** **(2020)** | To create speech emotion recognition system | An emotion recognition system based on speech signals in two-stage approach, namely feature extraction and classification engine. | The system involves a two-stage approach, first stage focusing on feature extraction using a combination of MFCC, ZCR, TEO, and HNR. The second involves dimension reduction using an auto-encoder and classification using a support vector machine (SVM). | It utilizes 39 Mel Frequency Cepstral Coefficients (MFCC), Zero Crossing Rate (ZCR), Harmonic to Noise Rate (HNR), and Teager Energy Operator (TEO) as audio features for emotion recognition. Additionally, an auto-encoder is employed for dimension reduction to select pertinent parameters from the extracted features. | The model achieved an accuracy rate of 74.07% for recognizing six emotions from the RML dataset. The results showed that the RBF kernel of SVM outperformed the linear and polynomial kernels in terms of recognition rates. | * **1.** Improved identification rates compared to other systems. * **2.** Utilization of auto-encoder for feature selection. | **1.** Limited dataset size  **2.** Empirical tuning of SVM parameters | The solution achieved a maximum identification rate of 74.07% for recognizing six emotions using the proposed feature extraction, dimension reduction with auto-encoder, and SVM classification. |
| **Husbaan I. Attar, Nilesh K. Kadole, Omkar G. Karanjekar, Devang R. Nagarkar, Prof. Sujeet (2022)** | The main objective is to improve a person’s speech fluency by accurately diagnosing stutter and suggesting eature training exercises for practice. | It eature of Voice Activity Detection to identify speech segments, Speech Segmentation for meaningful division, Signal Pre-Processing for conditioning audio, Feature Extraction extracting relevant speech eature, Emotion Classification utilizing machine learning, and Statistics Analysis of Emotion Frequency for insight. | The proposed speech emotion recognition system uses machine learning techniques to classify emotions expressed in continuous speech. The system includes feature extraction, feature selection, and classification stages to achieve high accuracy rates in real-time recording experiments. | The important features used in the proposed system include Log-Mel Spectrogram, Mel-Frequency Cepstral Coefficients (MFCCs), pitch, and energy. | The proposed system achieved high accuracy rates in emotion classification using eatur such as Long Short Term Memory (LSTM), Convolutional Neural Networks (CNNs), Hidden Markov Models (HMMs), and Deep Neural Networks (DNNs). | **1.** High accuracy rates achieved in real-time recording experiments.  **2.** It has potential applications in various fields. | **1.** Limited to the four emotions considered in the study.  **2.** It may not generalize well to other languages or eature. | The final result of the solution demonstrated effective speech emotion recognition with potential applications in fields such as mental state assessment in dangerous environments and customer satisfaction monitoring. |
| **Dr. Mrs. Gresha Bhatia, Binoy Saha, Mansi Khamkar, Ashish Chandwani , Reshma Khot (2021)** | The main objective is to improve a person’s speech fluency by accurately diagnosing stutter and suggesting eature training exercises for practice. | The components are stutter assessment, therapy suggestion, Gated Recurrent CNN (GRCNN) models, SVM model, and a mobile application. | The system uses deep learning models, Gated Recurrent CNN and SVM, to automate stutter assessment and personalize therapy recommendations. | Its main features include speech fluency improvement, personalized treatment, and continuous performance monitoring. | The solution achieves high validation accuracies of approximately 95% for identifying prolongation and 92% for repetition in speech audio. | **1.** Improved speech fluency.  **2.** It provides personalized treatment. | **1.** Potential overreliance on technology.  **2.** There is a huge requirement for robust data. | The results demonstrate the system’s effectiveness in accurately diagnosing stutter types and recommending appropriate therapies based on stutter descriptors and speech fluency improvement. |
| **Seema Badra (2019)** | The solution aims to automate the assessment of stuttering severity by proposing an approach that uses Automatic Speech Recognition (ASR), focusing on features like Mel frequency Cepstral Coefficients (MFCC) and employing Support Vector Machine (SVM) for classification. | The solution involves using Mel frequency Cepstral Coefficients (MFCC) for feature extraction and Support Vector Machine (SVM) for classification to automate the assessment of stuttering severity. | The solution involves an automatic recognition system that utilizes speech processing techniques to assess the severity of stuttering, aiming to reduce the manual workload of speech-language pathologists. | It incorporates feature extraction algorithms to analyze spectral and temporal features of speech, enabling the identification of repetitions and prolongations characteristic of stuttered speech. | The solution demonstrates promising performance in accurately recognizing and quantifying the rate of stuttering severity, showcasing its potential to enhance the efficiency of assessment processes. | **1.** It provides an automation of stuttering severity assessment there by reducing manual workload.  **2.** It provides comprehensive eature of speech patterns leading to more accurate assessments. | **1.** The need for rigorous validation and calibration for accuracy.  **2.** It has potential limitations in capturing nuanced aspects of stuttering. | The final results indicate a significant reduction in the time required for assessing stuttering severity, thereby improving theeature workflow of speech-language pathologists and potentially leading to more timely interventions for individuals with speech disorders. |
| **M. Mahendran, R. Visalakshi, S. Balaji (2021)** | The goal of the proposed solution is to detect dysarthria in patients using a CNN-based model that analyses various speech features. The problem that needs to be solved is the difficulty in diagnosing dysarthria, which is speech impairment caused by various underlying conditions. | A Convolutional Neural Network-based model to detect dysarthria in patients. The model analyses various speech eature such as zero crossing rates, MFCCs, spectral centroids, and spectral roll off. The TORGO speech signal database is used for training and testing the model. | The proposed solution utilizes a pairwise distance-based CNN to compare frame-level distance patterns between healthy and dysarthric speech representations, achieving high accuracy in dysarthric speech identification . | The main features of the solution include the use of phonetically-balanced AP representations, an end-to-end framework for feature extraction and classification, and generalizability across languages, surpassing state-of-the-art CNN-based systems in dysarthric speech identification | The proposed CNN-based model achieves an accuracy score of 93.87% in early dysarthric speech diagnosis, demonstrating promising results in detecting dysarthria | **1.** Generalizable across languages for dysarthria detection.  **2.** It provides end-to-end framework with optimized feature extraction, distance matrix computation, and classification. | **1.** Overfitting is a major concern.  **2.** Valid padding reduces the number of eature. | The final results show that the solution outperforms traditional CNN models and achieves high accuracy in dysarthric speech identification, indicating its potential for effective impairment management. |
| **Mustaqeem and Soonil Kwon(2020)** | To enhance the accuracy of a speech emotion recognition system. | The system eature of three stages: Audio signal processing, Deep stride convolutional neural network (DSCNN) and generating the final output. | The proposed system uses audio signal preprocessing to generate spectrograms, which are then fed into a DSCNN architecture to extract deep features and classify emotions. | Usage of a novel eature thresholding technique to remove noise and unimportant eature from speech. | The proposed DSCNN model demonstrates strong performance, achieving aneature accuracy of 81.75% on the testing dataset. | 1. Efficient training time  2.High Accuracy  3.Computational simplicity  4.Effective preprocessing | 1.Limited interpretability  2.Sensitivity to Hyperparameter Tuning  3.Generalization Challenges | The proposed model achieves an overall accuracy of 81.75% on the testing dataset, outperforming other state-of-the-art CNN model |
| **Mohan Ghai, Shamit Lal, Shivam Duggal, and Shrey Manik (2017)** | To recognize emotions in speech and classify them in different emotion output classes. | The components of the study include the emotional speech database, feature extraction, feature vector, and classifiers, which are used to recognize emotions in speech signals. | The system extracts feature from audio signals, such as energy and Mel Frequency Cepstral coefficients, and uses supervised learning algorithms to classify and recognize emotions in speech signals. | The system leverages perceptual features including Mel Frequency Cepstral coefficients (MFCC), energy of the speech signals, and the classification algorithms Support Vector Machine (SVM), Random Decision Forest, and Gradient Boosting. | The system achieved a maximum accuracy of 81.05% in recognizing emotions in speech signals, with the Random Decision Forest classifier providing the highest accuracy. | 1. High accuracy in recognizing emotions in speech signals.  2. Effective use of Mel Frequency Cepstral coefficients and energy of speech signals as feature inputs. | The system may misclassify happiness as anger and is susceptible to performance variations due to noise and emotional expression variability. | The proposed approach achieved high accuracy of 81.05% in recognizing emotions in speech signals, with Random Forest algorithm performing the best among the classifiers tested. |
| **Apeksha Aggarwal, Akshat Srivastava, Ajay Agarwal, Nidhi Chahal, Dilbag Singh, Abeer Ali Alnuaim, Aseel Alhadlaq, and Heung-No Lee (2022)** | To create and evaluate effective methods for Speech Emotion Recognition | The system comprises two-way feature extraction eatur, Principal Component Analysis (PCA), Deep Neural Network (DNN) and pre-trained VGG-16 model. | The system utilizes two different eatur of feature extraction to improve the effectiveness of emotion recognition. | Feature extraction plays a crucial role in the process of recognizing speech emotions. Utilizing PCA for feature extraction with DNN and utilizing pre-trained VGG-16 model for speech emotion recognition. | The performance of the proposed system is evaluated using multiple algorithms and two datasets. The RAVDESS dataset is found to provide significantly better accuracy than using numeric features on a DNN. | This model helps in dataset reduction, which may enhance model performance and generalization. | Limited Emotion classes, models are computationally expensive. | The proposed approach for the RAVDESS dataset, achieved an accuracy of 72%. For TESS dataset using VGG-16 model, it achieved accuracy of 90%. |
| **Tedd Kourkounakis, Amirhossein Hajavi, Ali Etemad(2020)** | The work goal of the paper is to address speech disfluencies and stutters in the workplace. | Feature extraction, recurrent layers, detection task, data, and annotation | FluentNet uses a combination of deep neural network techniques to detect speech disfluency. Additionally, FluentNet incorporates an attention mechanism to focus on the important parts of speech, allowing for better performance in detecting disfluencies. | This model incorporates a Squeeze-and-Excitation Residual Network (SE-ResNet) and bidirectional LSTM layers for effective stutter feature learning. | FluentNet achieves an average miss rate of 9.35% and an accuracy of 91.75%, surpassing other models and setting a new state-of-the-art. It outperforms previous models and benchmark models on the LibriStutter dataset as well. | The proposed model utilizes direct audio signals, spectrogram features and its innovative architecture providing an efficient framework for learning stutter specific features. | Limitations of the FluentNet includes challenges in classifying interjections, lack of sufficient eature data and poor performance on word repetitions and prolongations. | On the UCLASS dataset, FluentNet achieves an average miss rate of 9.35% and an accuracy of 91.75%, surpassing other models and setting a new state-of-the-art. |
| **Manas Jain, Shruthi Narayan, Pratibha Balaji, Bharath K P, Abhijit Bhowmick, Karthik R, and Rajesh Kumar Muthu (** | To identify speaker’s emotion. | The systems components include the input speech signal, feature extraction using MFCC and LPC, classification based on SVM, and the output. | The mechanism of the system involves the following steps:  1. Procuring the speech signal  2. Extracting features using MFCC and LPC  3. Using SVM to classify the features into four emotions: sadness, anger, fear, and happiness. | The features used in this system include MFCC (Mel-frequency cepstral coefficients), LPC (Linear Predictive Coding), pitch, energy, and speaker rate. These features are extracted from the input speech signal and used for emotion classification using the SVM algorithm. | The system achieved emotion classification using SVM with an accuracy of 85% on the UGA and LDC datasets, demonstrating its effectiveness in speech emotion recognition. | The advantages of the system include its ease of training, scalability to high-dimensional data, and the ability to handle non-linear classification tasks effectively. | The limitations may include the need to carefully select appropriate kernel functions for non-linear classification tasks and computational cost is high. | The system achieved decently good results in speech emotion recognition, with an accuracy of 85% using SVM on the UGA and LDC datasets for classification of emotions. |
| **Bagus Tris Atmaja, Akira Sasou. 2022** | To create a deep learning-based model, particularly based on a time-delay neural network (TDNN), capable of accurately identifying various types of stuttering disfluencies solely from acoustic signals | There are two components, feature extraction and Evaluation | StutterNet utilizes TDNN to process MFCC features, capturing temporal aspects for eature stuttering detection. | MFCC features extracted from speech signals, capturing stutter-specific characteristics for robust identification in StutterNet. | Proposed system provides good performance with accuracy: 74.07 | Acoustic signal reliance.  Few trainable parameters.  Promising stuttering detection results. | Limited research in stuttering.  Relatively small dataset size.  Context window optimization challenges. | The results indicate promising stuttering detection with StutterNet, outperforming a competitive method in various disfluency types. |
| **Shakeel A. Sheikh, Md Sahidullah . 2021** | The specific aim is to achieve accurate stuttering detection using a time-delay neural network (TDNN) architecture with Mel-frequency cepstral coefficients (MFCCs) | There are two components TDNN architecture, MFCCs | StutterNet uses TDNN to capture temporal context, enhancing acoustic features for eature stuttering detection. | Mel-frequency cepstral coefficients (MFCCs), capturing stuttering nuances, crucial for StutterNet’s acoustic analysis. | Proposed system provides good performance with accuracy: 80.01 | Accurate stuttering detection.  Efficient acoustic feature utilization.  Outperforms existing methods. | Limited evaluation on complex scenarios.  Reliance on a single dataset.  Optimization eature careful consideration. | The research paper demonstrates that StutterNet outperforms existing eatur, achieving a 4.69% gain in accuracy and a 0.03 increase in Matthew’s correlation coefficient (MCC). |
| **Colin Lea, Zifang Huang. 2023** | To address the challenges faced by people who stutter (PWS) in using speech recognition technology. | The components are: ASR models, dysfluency refinement, endpoint tuning | ASR tuning, endpoint adjustments, and dysfluency refinement; mechanisms improve speech recognition for people with stuttering. | Endpoint tuning, ASR model adjustments, and dysfluency refinement enhance speech recognition for individuals with stuttering. | Proposed system provides good performance with accuracy: 79.01 | Improved stuttered speech recognition.  Minimal data requirements.  Significant error rate reduction. | Some potential trade-offs.  User preferences may vary.  Limited context variability analysis. | What is the result of the research paper |
| **Phani Bhushan S, Vani H Y. 2021** | Stuttered Speech Recognition (SSR) system using Convolutional Neural Networks (CNN) and Weighted Mel Frequency Cepstral Coefficients. | The components are CNN , and Mel Frequency Cepstral Coefficients | Extracts features from stuttered speech using Weighted MFCC, classifies with CNN, enhancing speech recognition accuracy. | Weighted MFCC, CNN architecture, convolutional and pooling layers, achieving 92% accuracy in stuttered speech recognition. | Proposed system provides good performance with accuracy: 90 | Improved stuttered speech recognition.  Utilizes CNN for efficiency.  Achieves high accuracy rate. | Some potential trade-offs.  User preferences may vary.  Limited context variability analysis. | The research paper achieves a 92% accuracy rate in recognizing stuttered speech using the proposed Stuttered Speech Recognition (SSR) system with Convolutional Neural Networks (CNN) and Weighted Mel Frequency Cepstral Coefficients (MFCC). |
| **Suwon Shon , Pablo Brusco. 2021** | The primary goal of the research paper is to eature and demonstrate the effectiveness of leveraging pre-trained language models, particularly BERT, in the context of speech sentiment analysis. | ASR eatur, pre-trained language models (BERT), sentiment classifier, pseudo label-based semi-supervised training. | Embedding pre-trained language models in sentiment classifiers, enabling robust sentiment analysis in spoken language. | Two-step pipeline, end-to-end system, BERT-based models, pseudo label-based semi-supervised training, sentiment analysis metrics. | Proposed system provides good performance with accuracy: 81.02 | Enhanced sentiment analysis performance.  Reduced need for human eature.  Efficient knowledge transfer from text. | Limited discussion on ASR updates.  Dependency on pre-trained LMs.  Lack of exploration on other text corpora. | Embedding pre-trained language models in sentiment classifiers, enabling robust sentiment analysis in spoken language.The paper highlights the effectiveness of leveraging pre-trained language models, particularly BERT, in improving sentiment analysis performance |
| **Hadhami Aouani , Yassine Ben Ayed(2020)** | To create a speech emotion recognition system. | The system eature of two stages: feature extraction and classification engine. | This is an article about speech emotion recognition. It discusses using MFCC features and an autoencoder to improve the accuracy of speech emotion recognition. | It discusses two sets of features, MFCC and Teager Energy Operator. The Support Vector Machines (SVM) is used as a classifier method. | Proposed system with SVM Stacked AE 39MFCC, ZCR, TEO, HNR filtered by stacked AE shows accuracy of 74.07 | 1.Comprehensive Feature Utilization  2.Optimized Classification with Auto-Encoder | 1.Dependency on Feature Relevance  2.Sensitivity to Hyperparameter Tuning  3.Generalization Challenges  4.Complexity of Auto-Encoder Training  5.Limited Interpretability | The proposed emotion recognition system achieved recognition rates of 72.83% and 74.07% using a basic auto-encoder and stacked auto-encoder for feature dimension reduction, surpassing the system without feature selection (65.43%) on the RML emotion database. |
| **S. Lalitha, Shikha Tripathi and Deepa Gupta(2018)** | The paper aims to explore the effectiveness of perceptual-based speech featuresfor emotion detection using deep neural networks on the Berlin database. | The system comprises pre-processing, feature extraction (utilizing Mel, Bark, and inverted Mel filter banks, as well as additional eature), and classification using a deep neural network | The system involves preprocessing audio data, extracting perceptual features (MFCCs, PLPC, BFCC, RPLP, MFPLPC, IMFCC), and utilizing a deep neural network for effective emotion detection, achieving improved performance on the Berlin database. | The system leverages perceptual features including MFCCs, PLPC, BFCC, RPLP, MFPLPC, and IMFCC for enhanced emotion detection, optimizing recognition accuracy on the Berlin database. | The performance of the system, evaluated using deep neural networks with perceptual speech features, shows improved emotion recognition accuracy compared to conventional eatur, as demonstrated on the Berlin database. | The advantages of the proposed system include eature speech emotion detection accuracy through the utilization of perceptual-based features, such as Mel frequency cepstral coefficients (MFCCs), with DNNs | The limitations of the system may include potential challenges in handling diverse emotional expressions, dependency on the selected eature, and sensitivity to variations in the training dataset, affecting generalization to real-world scenarios. | The results of the system indicate improved emotion recognition accuracy, particularly in valence and arousal dimensions, using a combination of perceptual-based speech features and deep neural networks, as demonstrated on the Berlin emotion database. |
| **Sadeen Alharbi, Madina Hasan, Anthony Simons, and Shelagh Brumfitt(2017)** | To eature the feasibility of using machine learning for stuttering detection in children’s speech eature. | The system comprises three main components: pre-processing, speech feature extraction, and classification eatur, utilizing machine learning approaches such as HELM and CRF for stuttering event detection in children’s speech transcripts. | The system employs machine learning approaches, specifically HELM and CRF, to detect stuttering events in children’s speech transcripts, addressing challenges such as limited training data and high dimensionality. | The system utilizes features such as n-grams (up to 4-grams) and post-words extracted from normalized transcripts for detecting stuttering events, considering various stuttering types (sound repetitions, part-word repetitions, word repetitions, phrase repetitions, prolongations). | The performance of the proposed system is evaluated through machine learning approaches, HELM and CRF, for detecting stuttering events in transcripts of children’s speech, achieving improved results with CRF and data augmentation. | The advantages of the system include the application of machine learning techniques (HELM and CRF), successful detection of stuttering events in children’s speech transcripts, and eature performance with data augmentation, providing valuable insights for automated diagnosis. | Limitations include challenges related to the lack of eature data and the high dimensionality of the data, with a focus on stuttering event detection in children’s speech transcripts is highlighted. | Results indicate that CRF outperforms HELM by 2.2% in baseline experiments for stuttering event detection in children’s speech transcripts. Data augmentation proves beneficial, particularly for rarely eature events, contributing to improved system performance. |
| **Bassam Ali Al-Qatab and Mumtaz Begum Mustafa (2021)** | The work goal of this paper is to evaluate the effectiveness of acoustic features and feature selection methods in classifying dysarthric speech based on the severity of impairment, employing various classification algorithms, with a focus on assessing classification accuracy and ranking performance. | The system’s components encompass acoustic eature (prosody, spectral, cepstral, voice quality), feature selection eatur (Interaction Capping, Conditional Information Feature Extraction, Conditional Mutual Information Maximization, Double Input Symmetrical Relevance, Joint Mutual Information, Conditional Redundancy, Relief), and classification algorithms (Support Vector Machine, Linear Discriminant Analysis, Artificial Neural Network, Classification and Regression Tree, I Bayes, RF | The system operates by extracting acoustic features (prosody, spectral, cepstral, voice quality) from dysarthric speech, employing feature selection eatur (Interaction Capping, Conditional Information Feature Extraction, Conditional Mutual Information Maximization, Double Input Symmetrical Relevance, Joint Mutual Information, Conditional Redundancy, Relief), and utilizing classification algorithms (Support Vector Machine, Linear Discriminant Analysis, Artificial Neural Network, CART, I Bayes, Random Forest) to categorize speech severity levels. | The study investigates dysarthric speech features, encompassing prosody, spectral, cepstral, and voice quality, using seven feature selection eatur (Interaction Capping, Conditional Information Feature Extraction, Conditional Mutual Information Maximization, Double Input Symmetrical Relevance, Joint Mutual Information, Conditional Redundancy, Relief), and evaluates classification algorithms (Support Vector Machine, Linear Discriminant Analysis, Artificial Neural Network, Classification and Regression Tree, I Bayes, Random Forest) for severity level classification. | The classification accuracy of the dysarthric speech analysis ranges from 40.41% to 95.80%, utilizing acoustic features and feature selection eatur with six classification algorithms, including Support Vector Machine, Linear Discriminant Analysis, Artificial Neural Network, Classification and Regression Tree, I Bayes, and Random Forest. | Advantages include its potential as assistive technology for individuals with speech impairments, leveraging acoustic features and diverse feature selection methods, achieving classification accuracy ranging from 40.41% to 95.80% with various classification algorithms. | Limitations of the dysarthric speech classification system include potential challenges associated with data sparsity, both in language coverage and speech database size, impacting the effectiveness of the automatic speech recognition (ASR) system. | The dysarthric speech classification system achieved classification accuracy ranging from 40.41% to 95.80%, utilizing acoustic features and various feature selection eatur with six classification algorithms. |
| **Tedd Kourkounakis, Amirhossein Hajavi, Ali Etemad(2019)** | The goal of this work is to address the identification and classification of various forms of stuttering, focusing on acoustic features rather than language models. | The system’s components include a deep RNN for learning stutter-specific eature, bidirectional LSTM layers for sequential data analysis, and a classification module for identifying eature types of stuttering based on acoustic eature. | The system’s mechanism involves generating spectrogram feature vectors from audio clips, utilizing a deep residual neural network to extract stutter-specific features, employing bidirectional long short-term memory (LSTM) layers for sequential data analysis | The system employs spectrogram feature vectors, a deep residual neural network (ResNet) with convolutional blocks, and bidirectional LSTM layers with dropout mechanisms for effective identification and classification of various stutter disfluencies based solely on acoustic features. | The proposed system achieves an average miss rate of 10.03% in detecting and classifying different forms of stutter disfluencies, surpassing the state-of-the-art by almost 27%. | The system offers advantages such as robust detection and classification of various stutter disfluencies solely based on acoustic features, outperforming existing models with an average miss rate of 10.03%. | The limitations of the system include potential misclassification of longer utterances due to reliance on four-second windows, and challenges in accurately identifying certain stutter types, such as prolongation, leading to a slightly higher miss rate for these cases. | The proposed model achieved an average miss rate of 10.03%, outperforming the state-of-the-art by almost 27%, in the detection and classification of different forms of stutter disfluencies using acoustic features. |

**CHAPTER 3**

**PROPOSED SYSTEM**

* 1. **Proposed System**

In the proposed system, we introduce SpeechSentio, an AI-powered speech therapy platform incorporating both stutter detection and emotion analysis. Addressing limitations in traditional stuttering detection methods, SpeechSentio offers real-time analysis to pinpoint stuttering challenges and craft precise interventions. By integrating emotional intelligence, the platform creates a supportive therapy environment, reducing anxiety and enhancing motivation for individuals undergoing speech therapy. Personalized pronunciation practice is facilitated through in-depth analysis of individual speech patterns, optimizing practice time and accelerating progress toward fluency goals. The system’s architecture employs a dual-branch approach, combining preprocessing techniques with machine learning algorithms for both emotion recognition and stutter detection. Through this innovative approach, SpeechSentio aims to revolutionize speech therapy, offering tailored interventions and fostering meaningful interactions for improved treatment outcomes.

* 1. **Advantages of Proposed System**

The proposed system has the following advantages:

* Real-time Analysis: The system offers immediate insights into stuttering challenges, enabling prompt intervention and personalized therapy sessions.
* Enhanced Therapy Environment: By integrating emotion recognition technology, the system creates a supportive environment, fostering motivation and facilitating quicker progress in speech therapy.
* Personalized Practice: Through in-depth analysis of individual speech patterns, the system optimizes pronunciation practice, maximizing efficiency and accelerating progress towards fluency goals.
* Comprehensive Detection: Leveraging a dual-branch architecture and advanced preprocessing techniques, the system ensures thorough stutter detection and precise emotion analysis, enhancing diagnostic accuracy.
* Scalability: The system’s automated processes and machine learning algorithms enable scalability, allowing therapists to efficiently manage a large volume of cases and provide tailored interventions.
* Real-time Feedback: Therapists receive real-time feedback on therapy progress, enabling them to adjust interventions dynamically and optimize treatment plans for better outcomes.
* Improved Treatment Efficacy: By offering tailored interventions, real-time feedback, and precise analysis, the system ultimately improves the efficacy of speech therapy, leading to better treatment outcomes for individuals with speech impediments.
  1. **SYSTEM REQUIREMENTS**

The system requirements for the development and deployment of the project as an application are specified in this section. These requirements are not be confused with the end-user system requirements. There are no specific, end-user requirements as the intended application is cross-platform and is supposed to work on devices of all form-factors and configurations.

* + 1. **SOFTWARE REQUIREMENTS**

Below are the software requirements for application development:

1. Operating System: Compatible with Windows, macOS, or Linux distributions (e.g., Ubuntu).
2. Python Environment: Python 3.x installed with essential libraries such as TensorFlow, Keras, scikit-learn, and NumPy for machine learning and signal processing tasks.
3. Speech Recognition Libraries: Integration with libraries like PyAudio or SpeechRecognition for capturing and preprocessing speech input.
4. Development Tools: Integrated Development Environment (IDE) such as PyCharm or Jupyter Notebook for coding and experimentation.
5. Version Control: Git for version control management, facilitating collaboration and tracking changes in code and project files.
   * 1. **HARDWARE REQUIREMENTS**

Hardware requirements for application development are as follows:

1. Processor: Multi-core processor (e.g., Intel Core i5 or equivalent) for efficient computation.
2. Memory (RAM): Minimum 8 GB RAM to support data processing and machine learning tasks effectively.
3. Storage: SSD storage with sufficient capacity (at least 256 GB) for storing datasets, models, and application files.
4. Sound Input Device: High-quality microphone or headset for capturing speech samples with clarity.
5. Graphics Card: Optional dedicated GPU (e.g., NVIDIA GeForce GTX 1060) for accelerated deep learning computations.
   * 1. **IMPLEMENTATION TECHNOLOGIES**

* **TensorFlow**

TensorFlow is an open-source machine learning framework developed by Google, widely used for building and training deep learning models. It provides a comprehensive ecosystem of tools, libraries, and resources for various machine learning tasks, including natural language processing, image recognition, and speech analysis. TensorFlow offers flexibility in designing complex neural network architectures and efficient computation on both CPUs and GPUs, making it suitable for processing large-scale datasets and training sophisticated models for tasks such as emotion recognition and stutter detection in speech therapy applications.

* **Librosa**

Librosa is a Python package specifically designed for music and audio analysis. It offers functionality for extracting features from audio signals, including Mel-frequency cepstral coefficients (MFCCs), spectral features, and rhythm patterns, which are essential for tasks like speech processing and emotion recognition. Librosa provides an easy-to-use interface for loading audio files, manipulating signals, and extracting relevant features, making it a valuable tool for implementing speech analysis algorithms in research and practical applications.

* **Deep Learning**

Deep Learning is a subfield of machine learning that focuses on training neural networks with multiple layers (deep architectures) to learn intricate patterns and representations from data. Deep Learning models, such as Convolutional Neural Networks and Recurrent Neural Networks, have achieved remarkable success in various domains, including computer vision, natural language processing, and speech analysis. Deep Learning excels at capturing hierarchical features and capturing temporal dependencies in sequential data, making it well-suited for tasks like speech emotion recognition, where subtle nuances and context play a crucial role. Deep Learning frameworks like PyTorch and Keras provide high-level APIs for building, training, and deploying deep neural networks, facilitating the implementation of advanced speech therapy systems with state-of-the-art performance.

* **Ensemble Learning**

Ensemble learning is a machine learning technique where multiple models are combined to improve the overall predictive performance and generalization of the system. Instead of relying on a single model’s predictions, ensemble methods leverage the collective wisdom of multiple models to make more accurate and robust predictions. The key idea behind ensemble learning is that different models may capture different aspects of the data or learn different representations, leading to complementary strengths. Ensemble methods can take various forms, such as averaging predictions (e.g., bagging), combining predictions based on voting (e.g., Random Forest), or building a sequence of models where each subsequent model corrects the errors of the previous ones (e.g., boosting). By aggregating predictions from diverse models, ensemble learning can mitigate overfitting, reduce bias and variance, and improve the overall performance of the system. In speech therapy applications, ensemble learning can be used, for instance, to combine the predictions of multiple classifiers for stutter detection or emotion recognition, leading to more accurate and reliable results.

* **Sound File**

A sound file, also known as an audio file, is a digital representation of sound stored in a computer-readable format. Sound files typically contain sampled audio data captured from various sources, such as speech, music, or environmental sounds. Common audio file formats include WAV, MP3, FLAC, and AIFF, each with its own compression and encoding methods. Sound files can vary in duration, sampling rate, bit depth, and number of channels, depending on the recording equipment and settings used. In speech therapy applications, sound files are used to store recorded speech samples for analysis, processing, and diagnosis. These files serve as the input data for various signal processing techniques and machine learning algorithms aimed at tasks such as stutter detection, emotion recognition, and fluency analysis. Sound files play a crucial role in understanding and addressing speech disorders, facilitating personalized therapy interventions and monitoring progress over time.

* **K-Nearest Neighbors**

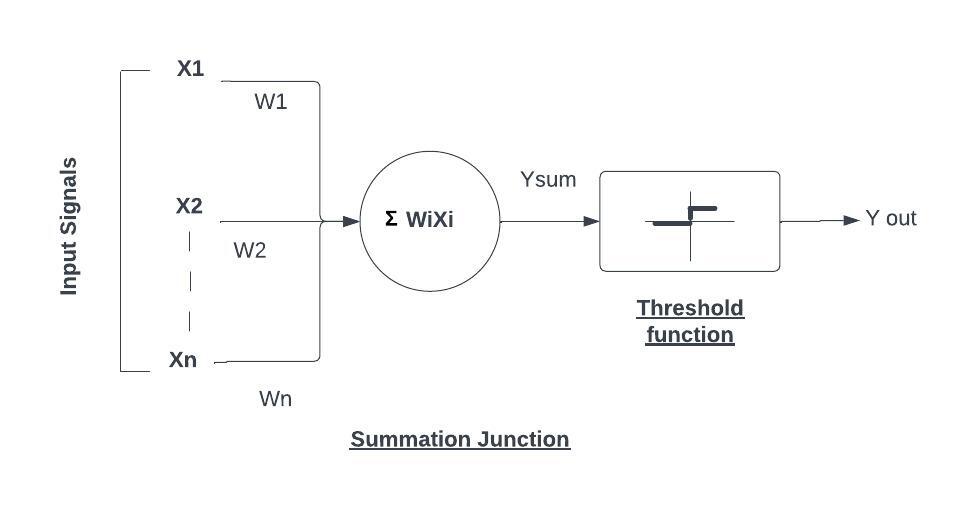
KNN is a simple yet effective machine learning algorithm used for classification and regression tasks. It operates on the principle of similarity, where the class or value of a new data point is determined by the majority vote or averaging of its K nearest neighbors in the feature space. KNN is non-parametric and instance-based, meaning it doesn’t make strong assumptions about the underlying data distribution and relies on the entire dataset for making predictions. In speech therapy applications, KNN can be utilized for tasks like stutter detection by analyzing similarities between speech features extracted from audio signals and those of known fluent or disfluent speech samples.

* **Decision Trees**

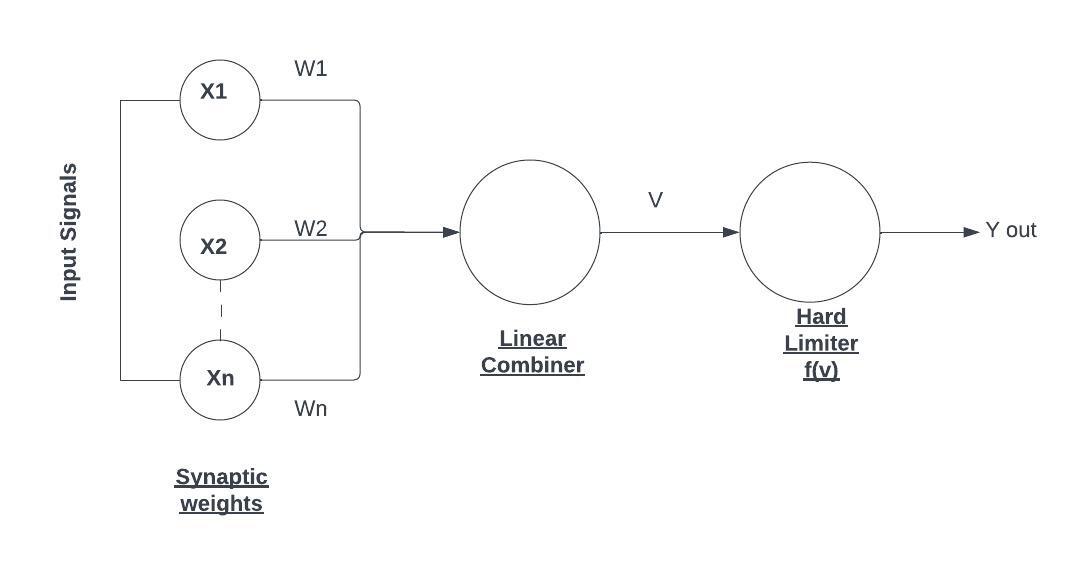
Decision trees are simple yet powerful machine learning models that recursively partition the feature space based on the value of input features to make predictions. Each internal node of the tree represents a decision based on a specific feature, and each leaf node corresponds to a predicted class or value. Decision trees are interpretable, easy to understand, and capable of handling both numerical and categorical data. In speech therapy applications, decision trees can be used for tasks like stutter detection by analyzing patterns in speech features extracted from audio signals. Additionally, decision trees can be part of ensemble methods like Random Forests, which combine multiple decision trees to improve performance and robustness in tasks such as stutter detection and emotion recognition.

* + 1. **Brief InTRODUCTION OF MLP**

Another essential element of the system’s architecture, and namely, the emotion recognition block, is the Multilayer Perceptron. The MLP is trained on an emotion-based labelled speech dataset and uses complex feature extraction methods special for speech processing, such as Mel-frequency cepstral coefficients, that allow analysis of spectral details of speech signals. As a result, due to the detected sophisticated connections between these features and a certain set of emotions, the MLP conducts efficient testing of how emotions can be identified in unseen speech samples. This novelty demonstrates the opportunity to test the usage of machine learning for emotion recognition tasks, demonstrating the system’s potential for enhancing speech therapy interventions with emotion analysis capabilities.



**Figure 1: McCulloch-Pitts Model**

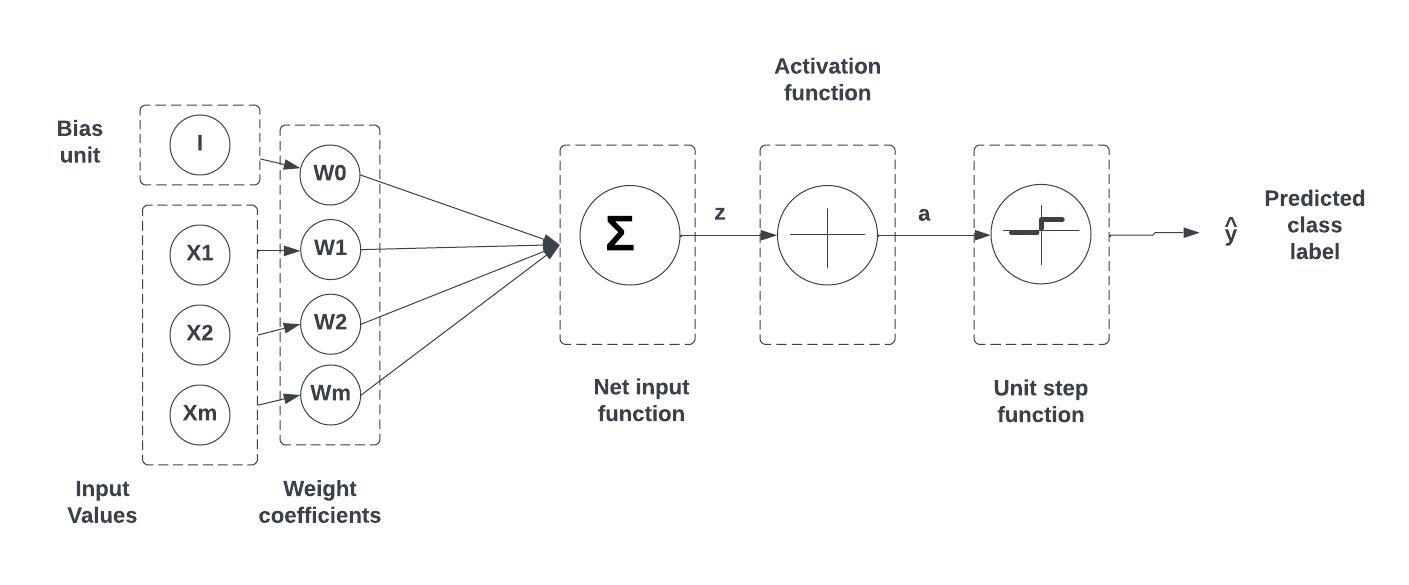


**Figure 2: Rosenblatt’s Perceptron**

The multilayer perceptron highlights a wealthy history dating back to the 1940s when Warren McCulloch and Walter Pitts displayed the concept of fake neurons, laying the establishment for neural organize inquire about. In any case, it wasn’t until the 1960s and 1970s that basic movements were made inside the outline of the perceptron, a single-layer neural system competent of twofold classification. In any case of the starting enthusiasm, the perceptron’s restrictions in dealing with nonlinear issues driven to a decay in intrigued.

The resurgence of intrigued in neural systems occurred within the 1980s with the ’ntroduction of the MLP, a more successful plan competent of learning complex designs through different layers of neurons. This breakthrough, coupled with the improvement of backpropagation, an compelling training methodology for altering network weights, revitalized the field of neural frameworks. Subsequently, MLPs have advanced with milestones in optimization methodologies, activation functions, and hardware capabilities, advancing to become a establishment of contemporary-era machine learning and laying the foundation for more progressed neural network structures.

Multilayer Perceptrons represent a foundational cornerstone in the domain of artificial neural networks, offering a versatile solution extensively applied across various machine learning and pattern recognition endeavours. The multi-layered architecture of the MLP, which can be described as the system of interconnected nodes of input, hidden, and output levels, boasts very significant capabilities in recognizing any intricate details and correlations present in the training datasets. The components of feed-forward multi-layer perceptrons are structured in such a way that they are able to transmit information smoothly from the input nodes through hidden layers to the output nodes that process tasks like classification, regression, or clustering.



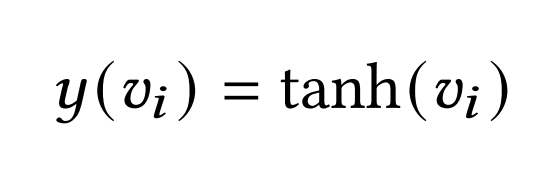
**Figure 3: Multi-layer Perceptron**

A multi-layer perceptron is a computational device in the sense that the nodes mitigate to the machines by doing their transformation functions, giving weights to input data, and sending outputs to layers above. The hidden layers feature map the structures in complex data distribution spaces, the model can be enhanced through non-linear interactions and higher order features being captured. BP helps neural networks in learning faster during the training process by acting as a learning signal that drives changes in the model’s input biases and weights. Such changes are fine-tuned in the direction of a shrinking gap in the differences between the actual outputs and the expected ones. Thus, the model parameters become optimized.

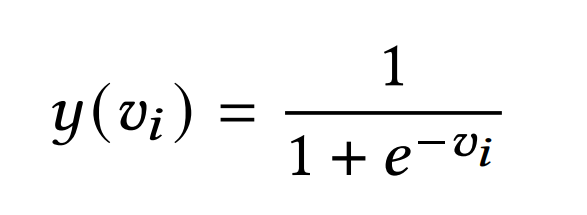
That is a nonlinear chain of activities, namely predictor generation, error computation, and in trajectory of update, for the parameters, which finally yields introduction of predictive capabilities for the MLP. Among a vast ensemble of neural network architectures, MLPs have been highly praised for their performance handling structured data and adaptability for representing non-linear linkages which extends its applicability in a broad range of domains, such as speech recognition and image classification.

**Activation Function:** When all neurons within a multilayer perceptron employ a linear activation function, meaning a function that directly maps the weighted inputs to each neuron’s output, linear algebra demonstrates the reduction of any number of layers to a simplified two-layer input-output model. However, in MLP’s various neurons utilize nonlinear activation functions, which were designed to emulate the firing frequency of biological neurons’ action potentials.

The activation functions that have been historically common are’ both sigmoid, and they can be described as:



and



The first is a hyperbolic tangent that ranges from -1 to 1, while the other is the logistic function, which is similar in shape but ranges from 0 to 1. Here yi represents the output of the ith node (neuron), and vi is the weighted sum of the input connections.In the realm of alternative activation functions, several proposals have emerged, encompassing options such as the rectifier and softplus functions. Additionally, more specialized activation functions, like radial basis functions utilized in radial basis networks, have been proposed, representing another class of supervised neural network models.Notably, recent advancements in deep learning have seen the widespread adoption of the rectified linear unit (ReLU) as a prominent solution to address numerical challenges associated with traditional sigmoid functions.

**CHAPTER 4**

**SYSTEM DESIGN**

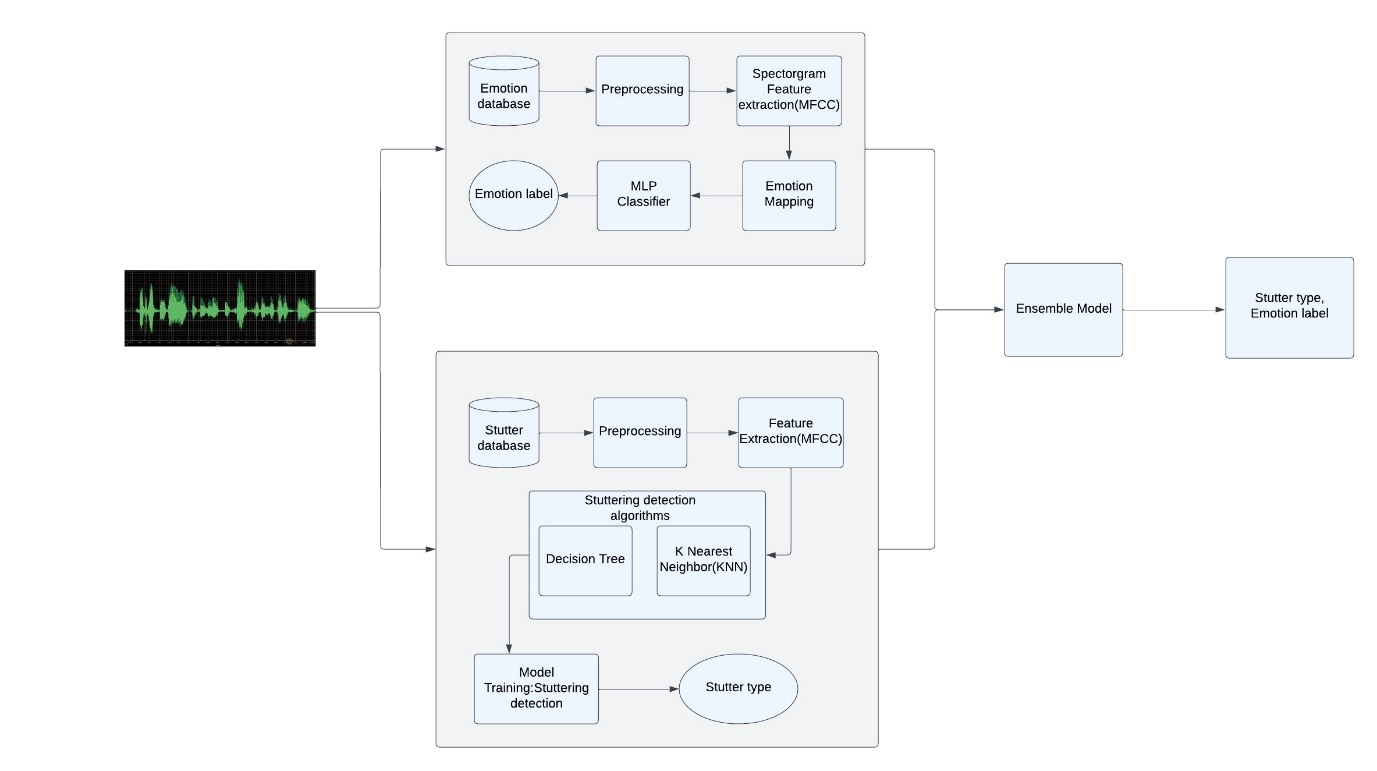
**4.1 PROPOSED SYSTEM ARCHITECTURE**

The architecture of the proposed system for AI-powered speech therapy with emotion analysis involves a dual-branch approach that combines speech recognition with emotion analysis and stutter detection. The system begins with preprocessing steps to prepare the raw speech signal for analysis. These steps include noise reduction, silence removal, and normalization to ensure optimal feature extraction.

In the emotion branch of the architecture, a spectrogram is generated, and Mel-frequency and MFC features are extracted. These features are then fed into a MLP classifier trained to recognize emotions based on speech patterns. The MLP leverages advanced feature extraction techniques, such as MFCCs, to analyze spectral variations in speech signals and classify emotions in unseen speech samples

On the other hand, the stutter detection branch also utilizes MFCC extraction but employs algorithms like decision trees or KNN to analyze temporal variations within the MFCC sequence. These temporal variations can indicate disfluencies like stutters, where the speaker might hesitate, repeat sounds, or prolong syllables. By analyzing these patterns, the algorithms can potentially detect the presence of stutters within the speech signal.

An optional ensemble model may be considered to combine these algorithms for potentially improved stutter detection accuracy. This architecture allows for the recognition of emotions and analysis of stuttering patterns simultaneously, providing a comprehensive understanding of the emotional aspects associated with stuttering behavior. By integrating emotion analysis with stutter detection, the system aims to enhance the effectiveness of therapeutic interventions and tailor treatment plans to individual needs in the field of speech therapy. In the emotion branch of the architecture, a spectrogram is generated, and MFCC features are extracted. These features are then fed into a MLP classifier trained to recognize emotions based on speech patterns. The MLP leverages advanced feature extraction techniques, such as MFCCs, to analyze spectral variations in speech signals and classify emotions in unseen speech samples.



**Figure 4: Architecture of proposed system**

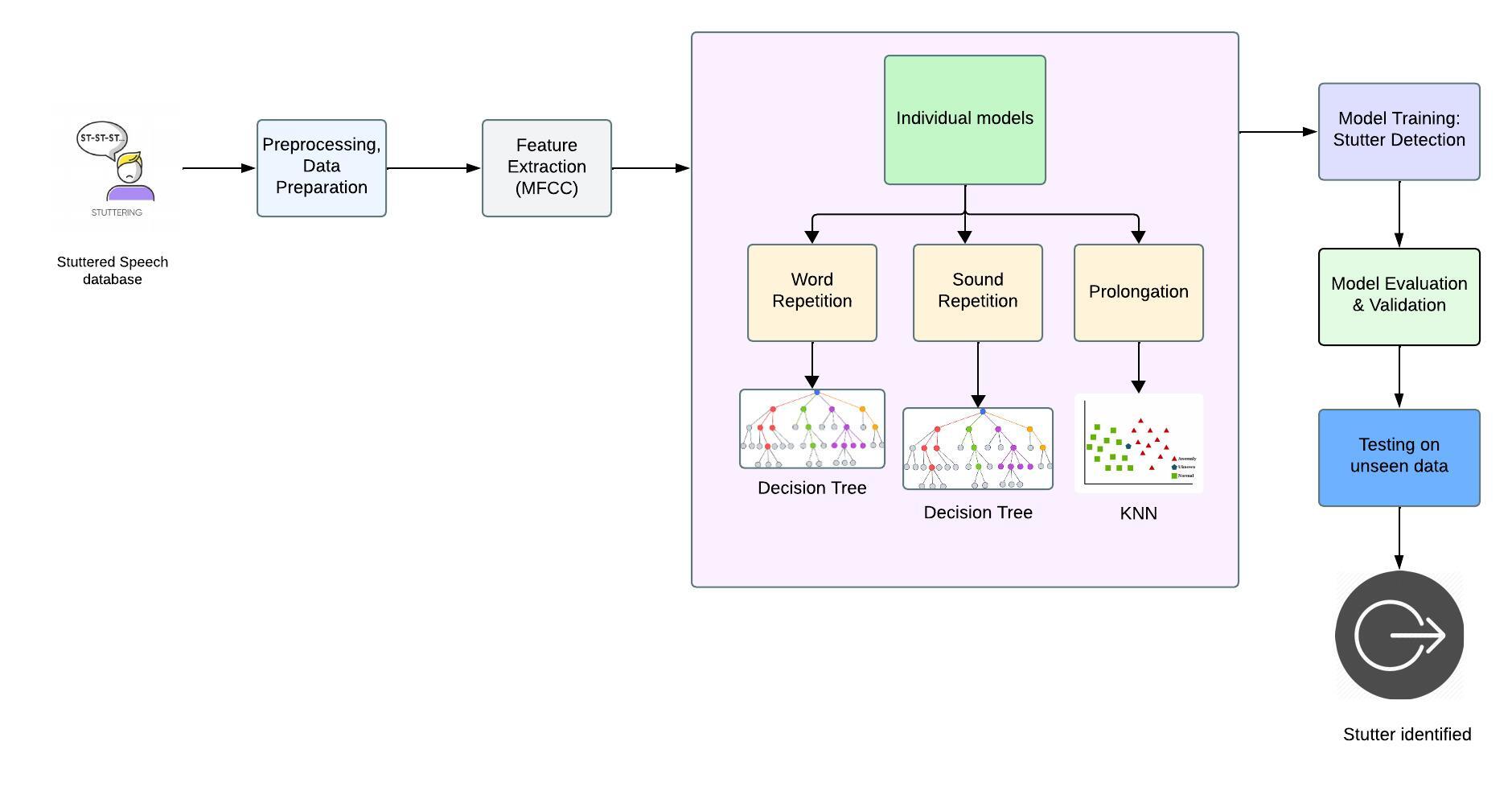
An optional ensemble model may be considered to combine these algorithms for potentially improved stutter detection accuracy. This architecture allows for the recognition of emotions and analysis of stuttering patterns simultaneously, providing a comprehensive understanding of the emotional aspects associated with stuttering behavior. By integrating emotion analysis with stutter detection, the system aims to enhance the effectiveness of therapeutic interventions and tailor treatment plans to individual needs in the field of speech therapy.

**4.1.1 stutter identification**

The provided image outlines a system architecture designed for detecting speech stuttering using machine learning techniques. Here is a concise overview of the process:

**1.Data Preparation:** The system gathers speech data containing instances of stuttered speech from various sources such as clinical records, public databases, or recorded samples

**2.Preprocessing:** Raw speech data undergoes preprocessing to make it suitable for machine learning analysis. MFCC’s are utilized to extract spectral features from the speech signals, facilitating the distinction between fluent and stuttered speech segments.

**Figure 5: Workflow of Stutter detection process**

**3.Model Training:** The system employs two main models:

* Speech Recognition Model: Converts speech audio into textual transcripts, enabling further analysis.
* Stutter Detection Model: Classifies speech segments as fluent or stuttered. This model utilizes the extracted MFCC features for accurate classification.

**4.Model Evaluation and Validation:** The performance of the models is evaluated using relevant metrics, focusing on characteristics like word repetitions and sound prolongations indicative of stuttered speech. Additionally, the models are validated on unseen data to ensure their effectiveness across diverse speech samples.

**5.Decision Tree:** Following model evaluation, a decision tree is utilized to make final classifications. Decision trees employ a hierarchical structure to classify data, with splits based on decision rules derived from the models' outputs. This aids in determining whether a speech segment is fluent or stuttered, based on the combined insights from the speech recognition and stutter detection models.

After undergoing a meticulous process encompassing data preparation, preprocessing, model training, and evaluation, the system provides crucial insights for stutter detection. These insights manifest as classification results or decision outputs, representing the system's analysis of the input speech data. Acting as actionable information, these outputs empower users to discern stuttered speech segments from fluent ones, aiding in decision-making and intervention strategies. By leveraging advanced machine learning methodologies and neural network architectures, the system equips users with robust tools to navigate the complexities of speech analysis. Through these advancements, the system fosters a deeper comprehension of speech disorders and facilitates effective interventions, ultimately contributing to improved communication and quality of life for individuals affected by stuttering.

| **Stutter Type** | **Description** |
| --- | --- |
| Word Repetition | The repetition of whole words or phrases within speech, often resulting in the same word being repeated multiple times. |
| Sound Repetition | The repetition of specific sounds or syllables within words, causing a stuttering effect within the pronunciation. |
| Prolongation | The elongation of sounds or syllables within words, leading to a drawn-out speech pattern or hesitation in speech. |

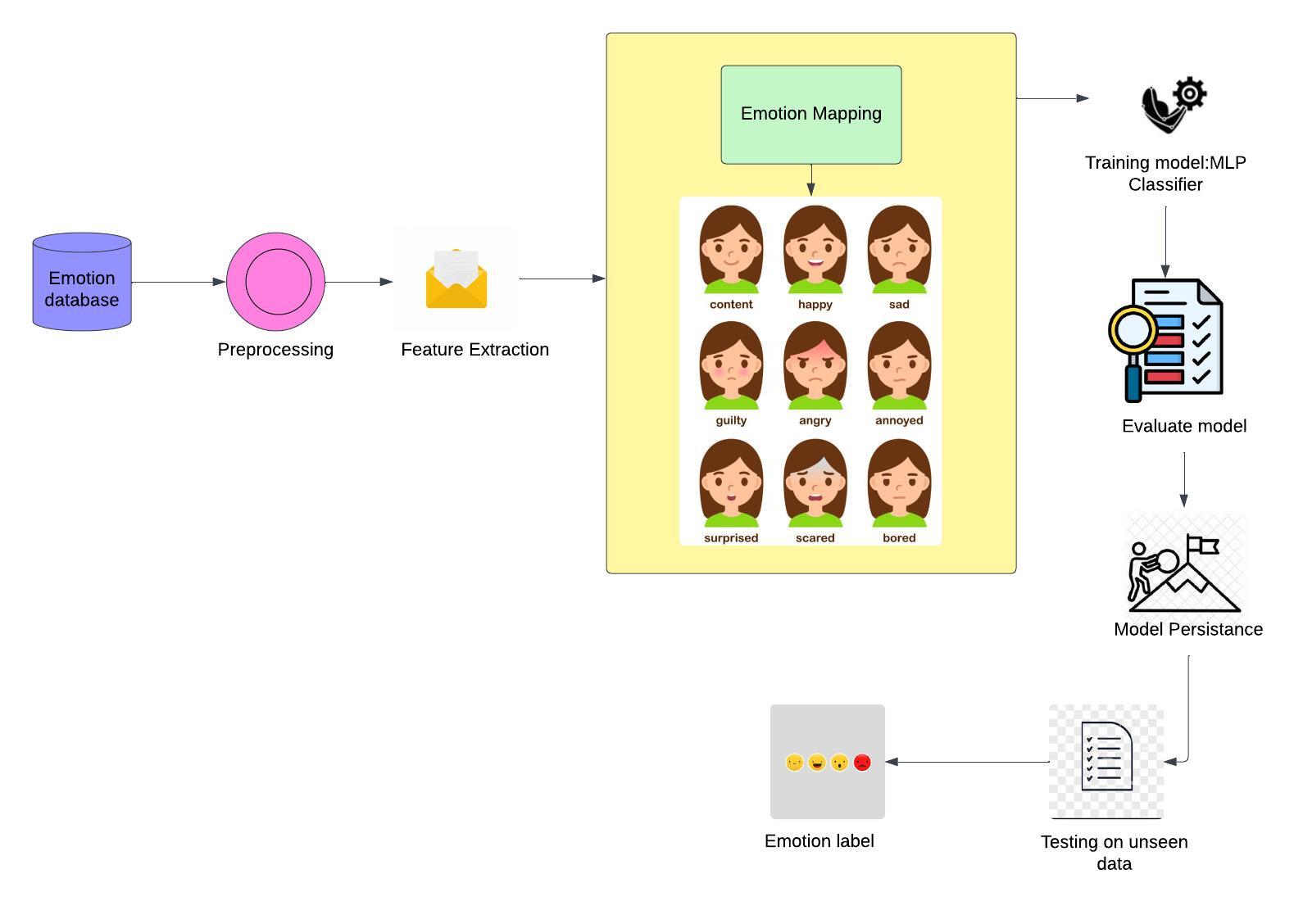
**Table 4.1.1 Types of stutters and Descriptions**

**4.1.2 Emotion recognition**

The architecture for an emotion detection system involves a systematic approach to analyze various forms of content, such as text, audio, or video, to identify underlying emotional states. Here's an overview of the key stages:

**1.Data Acquisition:** This initial phase involves sourcing diverse content capable of eliciting emotional responses. This content may span various mediums such as textual articles, audio recordings, or video clips, each potentially evoking different emotional states.

**2.Preprocessing:** Raw content undergoes preprocessing to ensure uniformity and prepare it for subsequent analysis. Textual data may undergo normalization processes, converting uppercase letters to lowercase, removing punctuation, and reducing words to their root forms through corresponding emotions. This process enables the model to recognize patterns and associations between features and emotions.



**Figure 6: Workflow of Emotion Analysis process**

**3.Feature Extraction:** Following preprocessing, specific features are extracted from the content to identify emotional cues effectively. Techniques vary depending on the nature of the content:

* Text Analysis: This involves sentiment analysis to gauge the emotional tone of the text, as well as extracting keywords or phrases indicative of specific emotions.
* Audio Analysis: For audio content, features like pitch, loudness, and spectral characteristics are extracted to discern emotional nuances in speech or sound.

**4.Model Training:** A MLP neural network, a type of artificial neural network suited for classification tasks, is employed for training. The MLP learns from a labeled dataset of emotional content, adjusting its internal parameters iteratively to accurately map extracted features to corresponding emotions. This process enables the model to recognize patterns and associations between features and emotions.

**5.Model Classification:** Once trained, the MLP classifier is utilized to predict emotions in new, unseen content. Extracted features from the content are fed into the classifier, which generates an output indicating the predicted emotion based on the learned patterns from the training phase.

**6.Model Persistence:** Following training, the MLP model is serialized and stored for future use through techniques like model persistence. This facilitates efficient loading of the trained model without the need for repeated training on the entire dataset, enhancing system performance and resource utilization.

**7.Evaluation:** The system's performance is rigorously assessed using various metrics relevant to emotion detection. This evaluation phase ensures the model's accuracy and effectiveness in classifying emotions across different types of content.

**8.Testing on Unseen Data:** To validate its generalizability and robustness, the trained model is tested on new, unseen data not encountered during the training phase. This step is critical for assessing the model's ability to accurately detect emotions in real-world scenarios.

**9.Output:** The ultimate output of the system is an emotion label corresponding to the detected emotional state within the analyzed content, providing valuable insights into the emotional context of the data.

Following a rigorous process encompassing data acquisition, preprocessing, feature extraction, model training, and evaluation, the system delivers insightful outputs. These outputs, in the form of emotion labels, represent the detected emotional states within the analyzed content, offering valuable insights into its emotional context. Serving as a bridge between raw content and actionable insights, these labels empower users to interpret emotional nuances in textual, auditory, or visual content, enabling informed decision-making and personalized user experiences tailored to individual emotional preferences and responses. Leveraging advanced machine learning techniques and neural network architectures, the system equips users with powerful tools to navigate the complex landscape of human emotions in digital communication, fostering a deeper understanding of emotions in the digital age.

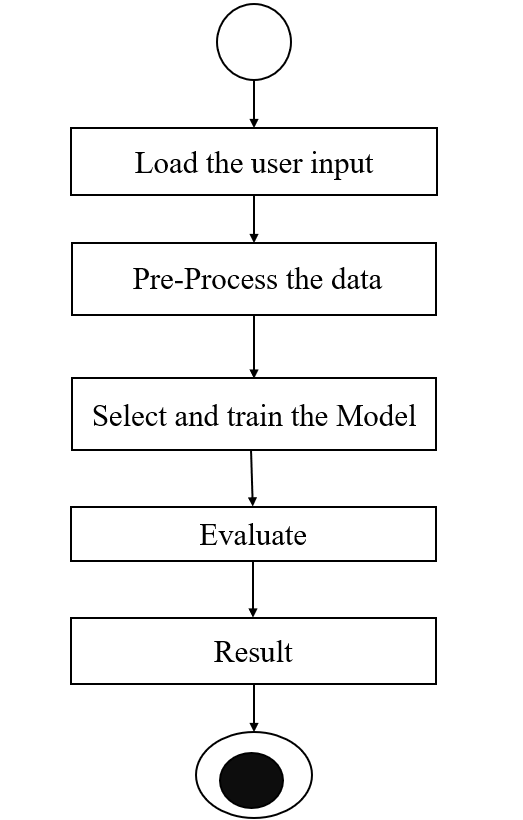
**4.2 UML diagrams**

Unified Modeling Language diagrams are graphical tools used in software engineering to visually represent various aspects of a system's structure and behavior. They include class diagrams for illustrating class relationships, use case diagrams for depicting user interactions, sequence diagrams for showing object interactions, and others like activity, state, component, and deployment diagrams. UML diagrams aid in communication among stakeholders, facilitating system understanding, design, and documentation throughout the software development process.

**4.2.1 Use Case diagram**

Use case diagrams in Unified Modeling Language visually represent the interactions between system components and external users or actors, showcasing different use cases and their relationships

The Use Case diagram shows the steps involved in speech stutter classification. It starts with loading the user's speech input, then preprocesses the data, selects, and trains a model, evaluates the model's performance, and finally displays the results.

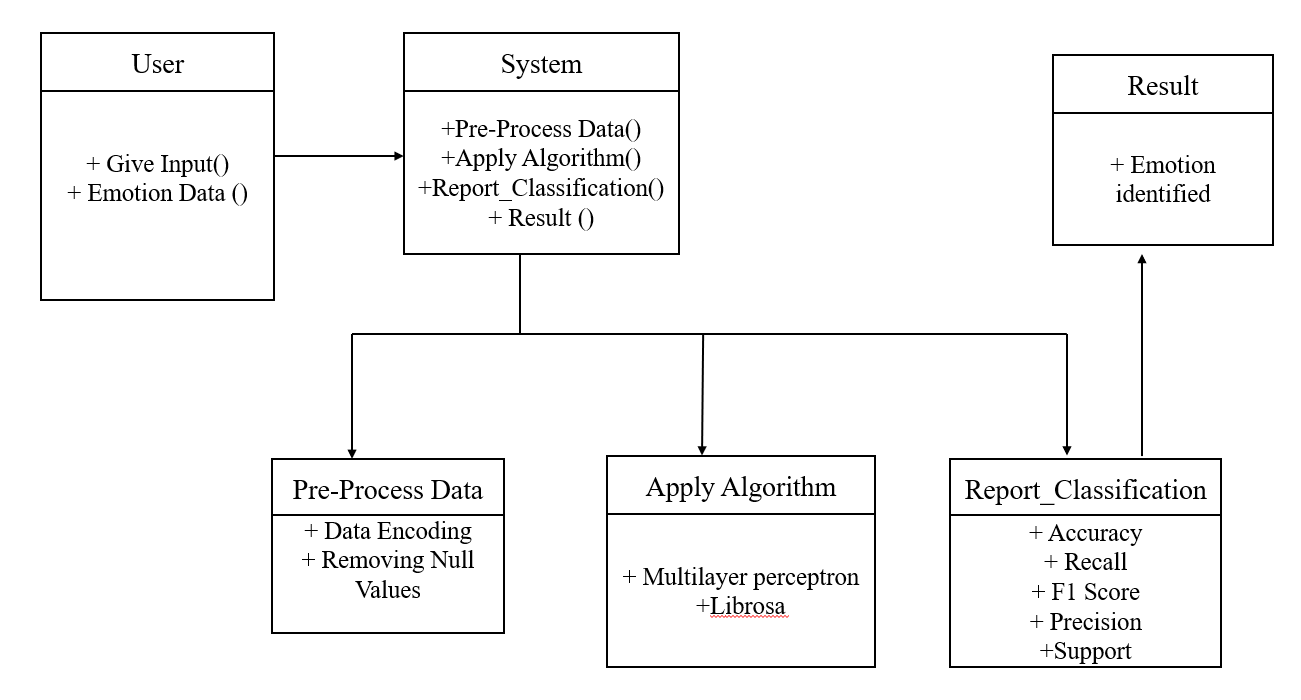
****

**Figure 7: Speechsentio System Use Case Diagram**

**4.2.2 Class Diagram**

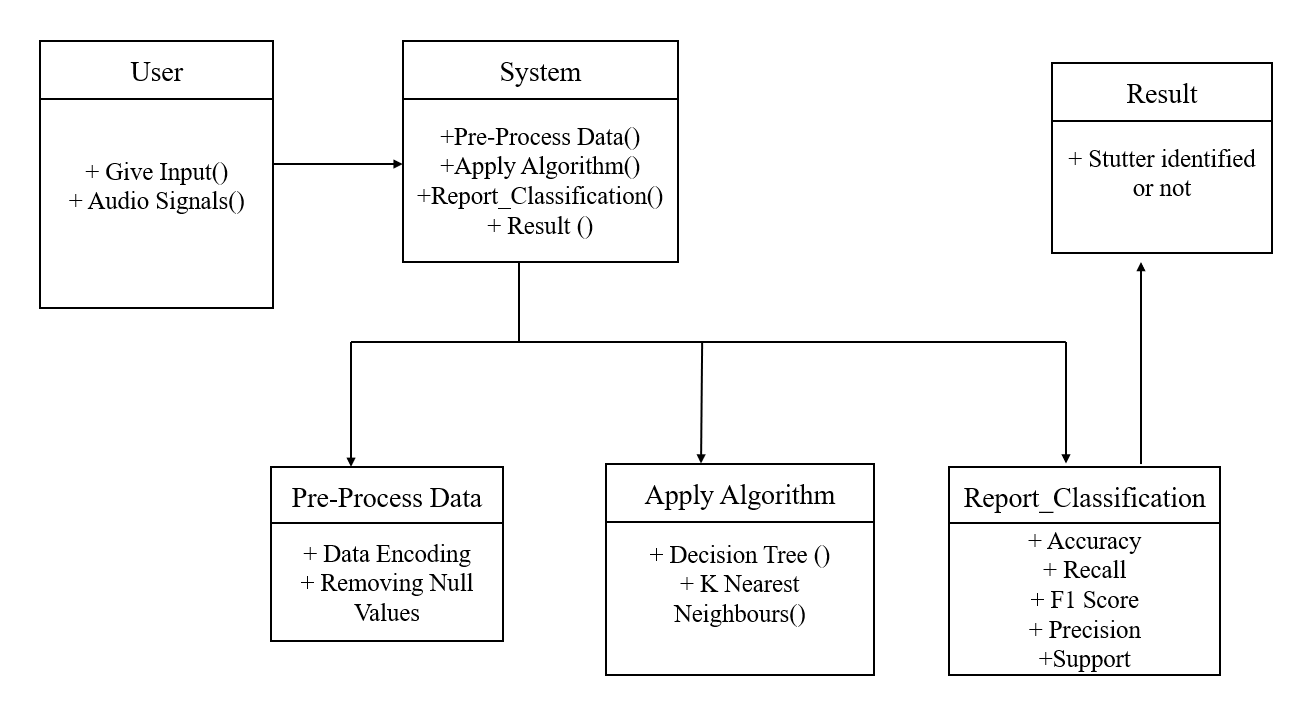
Class diagrams in Unified Modeling Language provide a visual representation of the structure of a system by illustrating classes, their attributes, methods, and the relationships between them. They serve as a blueprint for software design, depicting the static aspects of the system's architecture. Class diagrams facilitate communication among stakeholders, aid in system understanding, and guide developers in implementing the software according to the specified design.

The diagram depicts a system designed to analyze user input, likely emotion data. After preprocessing the data to ensure its quality, the system employs an algorithm to categorize the emotions. The system then generates a report evaluating the algorithm's performance using metrics like accuracy.

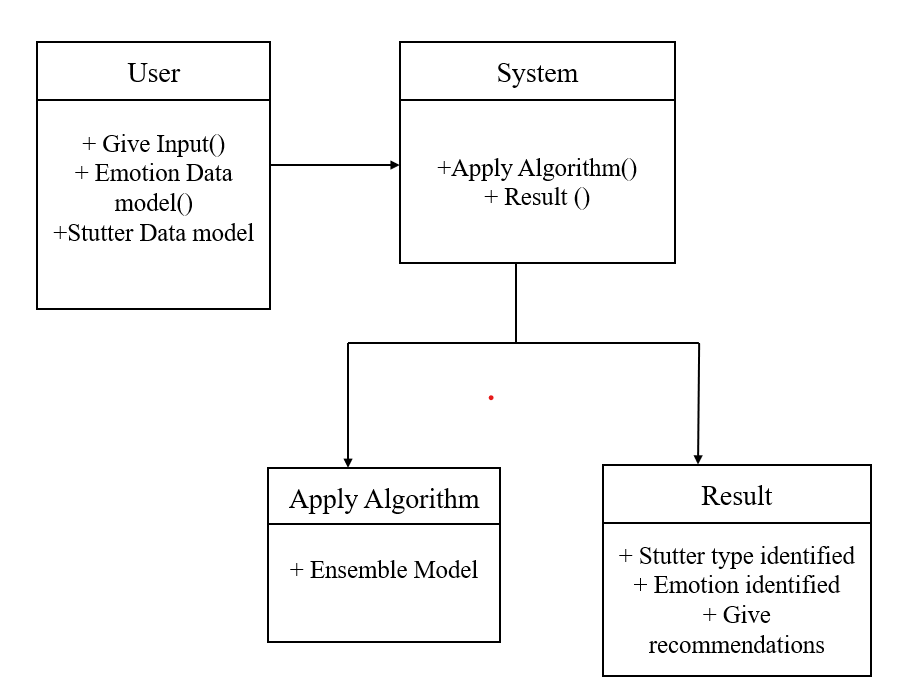


**Figure 8: Class diagram for Emotion Analysis model**

This above diagram outlines a stutter detection system. Users input audio data, which is then preprocessed and analyzed using two algorithms: a decision tree and K Nearest Neighbors. Finally, the system generates a report indicating the presence or absence of stuttered speech. While the diagram mentions performance metrics, it doesn't show how they're calculated.



**Figure 9: Class diagram for Stutter Detection model**



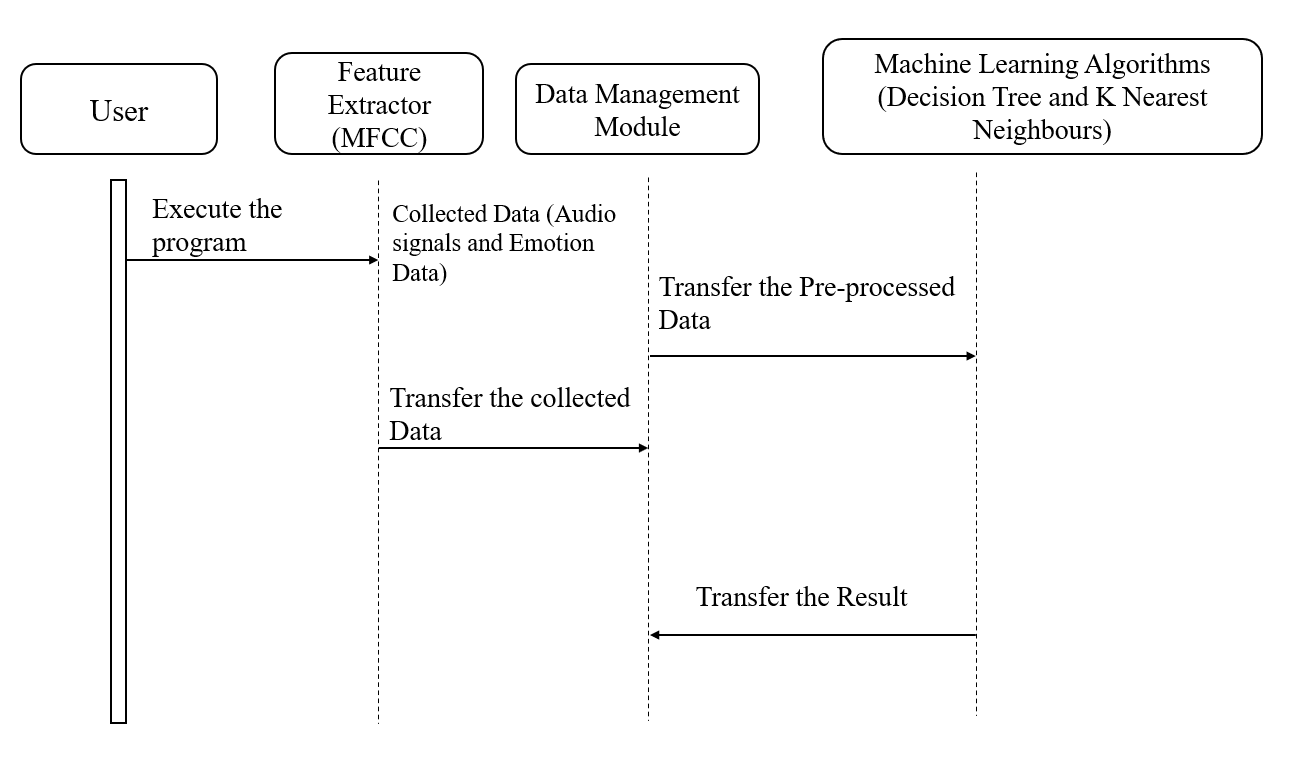
**Figure 10: Class diagram of a system that uses emotion data to identify a stutter type**

The diagram depicts systems for analyzing user input, likely emotion data, and detecting stuttered speech. Preprocessing is followed by algorithmic analysis, with resulting reports generated. The class diagram illustrates interactions between User, System, and StutterData classes, with User inputting data, System processing it, and results returned to the User class.

**4.2.3 Sequence diagram**

The sequence diagram illustrates the flow of interactions between system components during the execution of a specific scenario.

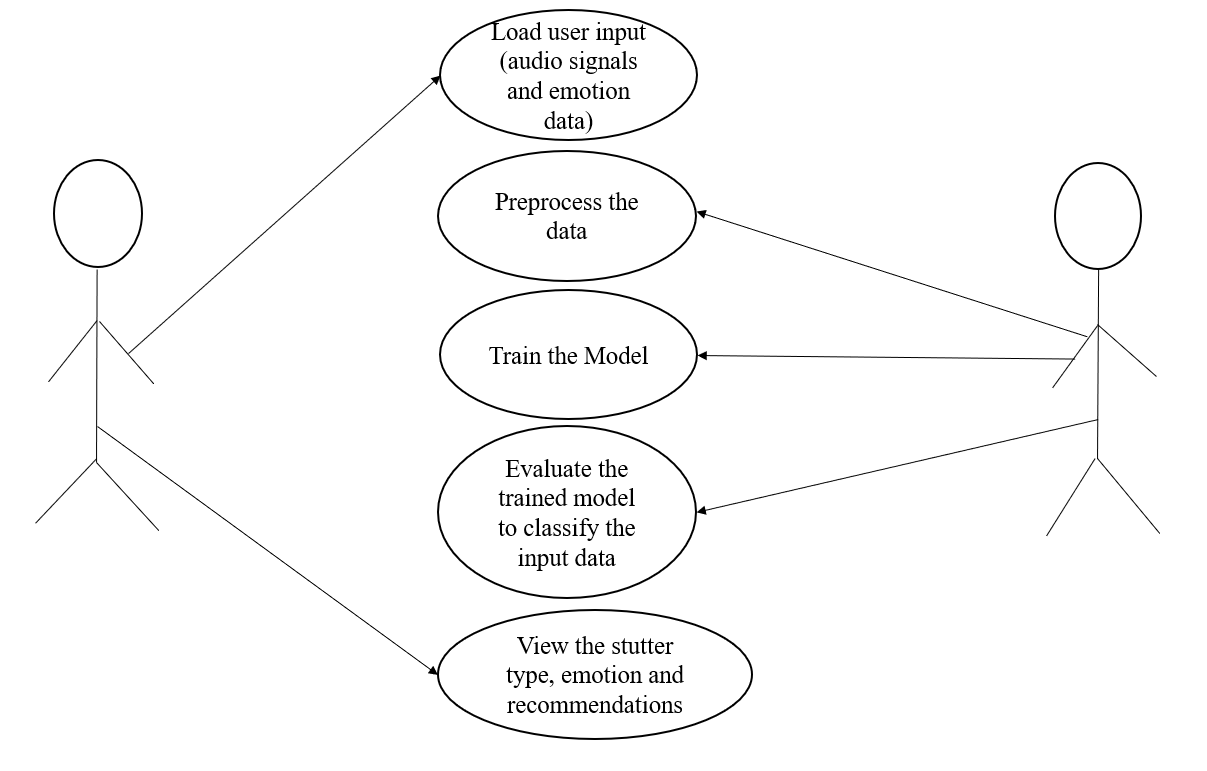
This sequence diagram illustrates the flow of a speech stutter classification system. It starts with the user providing speech input. The system then extracts features from the audio, classifies the stutter type using machine learning algorithms, and finally presents the classification results to the user.

****

**Figure 11: Sequence diagram of Speechsentio**

**4.2.4 Activity Diagram**

The activity diagram depicts the process of training a machine learning model for speech stutter classification. It starts with loading user input (audio and emotion data), followed by preprocessing the data. Then, the model is trained. Once trained, the model is evaluated to assess its performance in classifying stutters. Finally, the results are viewed, including the stutter type, emotion, and recommendations.

****

**Figure 12: Activity Diagram of Speechsentio**

**CHAPTER 5**

**IMPLEMENTATION**

**5.1 A BRIEF INTRODUCTION OF IMPLEMENTATION**

The implementation aims to develop a comprehensive system for both stutter classification and speech emotion recognition, utilizing machine learning algorithms. Stutter classification involves analyzing speech fluency to identify instances of stuttering, while speech emotion recognition focuses on detecting the emotional content conveyed in speech. The implementation process comprises three main subsections: stutter classification, speech emotion recognition, and overall execution. In the stutter classification section, data preprocessing techniques are applied to prepare the dataset, followed by building machine learning models such as Decision Trees and K-Nearest Neighbors (KNN) for classification. Similarly, in the speech emotion recognition section, data preprocessing is conducted to extract relevant features like Mel-frequency cepstral coefficients (MFCCs), followed by training models like Multi-Layer Perceptron (MLP) for emotion recognition. Finally, the overall execution involves integrating the stutter classification and speech emotion recognition modules into a unified system, evaluating its performance, optimizing for better results, and deploying the solution for real-world applications. This implementation aims to address both stutter classification and speech emotion recognition tasks, contributing to advancements in speech analysis and human-computer interaction domains.

**5.1.1 STUTTER CLASSIFICATION**

1. **Data Preprocessing:**

In the stutter classification process, the initial phase revolves around data preprocessing. Here, the primary objective is to curate a comprehensive dataset containing speech samples annotated with relevant information regarding stuttering occurrences. This dataset is often sourced from reputable repositories like Kaggle or specialized research databases, ensuring its quality and diversity. Subsequently, the collected data undergoes a series of preprocessing steps aimed at refining the speech samples and standardizing their format. Techniques like noise reduction, silence removal, and normalization are applied to enhance the quality and consistency of the speech data. Additionally, advanced signal processing methods may be employed to extract discriminative features such as Mel-frequency cepstral coefficients, which effectively encapsulate the distinctive characteristics of the audio signals, aiding in the subsequent classification process.

1. **Building Models with ML Algorithms:**

Following data preprocessing, the focus shifts towards building machine learning models tailored for stutter classification. Among the myriad of available algorithms, Decision Trees and KNN are prominent choices due to their simplicity and effectiveness in handling classification tasks. Decision Trees operate by recursively partitioning the feature space based on feature splits, effectively creating a hierarchical structure that facilitates sample classification. Conversely, KNN relies on the principle of similarity, where samples are classified based on the majority class of their nearest neighbors in the feature space. These algorithms are initialized with suitable parameters and trained on the preprocessed training data to learn the underlying patterns indicative of stuttering speech. Subsequently, the performance of the trained models is rigorously evaluated using various metrics such as accuracy, precision, recall, and the confusion matrix. This comprehensive assessment provides valuable insights into the effectiveness and robustness of the classification models, guiding further refinements and optimizations to enhance their classification capabilities. Through this iterative process of data preprocessing and model building, the stutter classification subsystem endeavors to provide an accurate and reliable framework for identifying instances of stuttering within speech recordings, thereby facilitating early detection and intervention for individuals with speech disorders.

**5.1.2 SPEECH EMOTION RECOGNITION**

1. **Data Preprocessing:**

In the domain of speech emotion recognition, the preparatory phase commences with meticulous data preprocessing, a pivotal step aimed at refining raw audio inputs into a structured format conducive to subsequent model training and evaluation. This process begins with the acquisition of suitable datasets curated specifically for emotional speech analysis, with prominent choices including repositories like the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS), revered for its diverse collection of annotated emotional speech samples. Upon dataset acquisition, the raw audio recordings undergo a series of preprocessing steps designed to extract salient features essential for discerning emotional states. Techniques such as feature extraction, prominently including Mel-frequency cepstral coefficients, chroma, and mel spectrogram, are employed to capture the spectral and temporal characteristics inherent in emotional speech expressions. Additionally, preprocessing may entail data augmentation strategies to augment the training dataset's diversity and improve model robustness. Once the preprocessing pipeline is complete, the refined dataset is partitioned into distinct training and testing subsets, ensuring the integrity of the model evaluation process and mitigating the risk of overfitting.

1. **Building Models with ML Algorithms:**

Following the comprehensive preprocessing of the emotional speech dataset, the subsequent phase entails the development and training of machine learning models tailored for speech emotion recognition tasks. Among the myriad of classification algorithms, the MLP stands as a prevalent choice owing to its capacity to learn intricate patterns in high-dimensional data. The model initialization process involves specifying the network architecture, including the number of hidden layers, neurons per layer, and activation functions, tailored to the intricacies of the emotional speech recognition task. Subsequently, the initialized MLP model is trained on the preprocessed training dataset using optimization techniques like SGD or Adam optimization. During the training phase, the model iteratively refines its parameters by minimizing a predefined loss function, thereby learning to map input speech features to corresponding emotional labels. Following convergence, the trained MLP model is subjected to rigorous evaluation using a separate testing dataset, where its performance is assessed across various metrics such as accuracy, precision, recall, and F1-score. This meticulous evaluation process serves as a litmus test for the model's efficacy in accurately discerning emotional states from unseen speech samples, thereby validating its utility in real-world applications.

**5.1.3 DEPLOYMENT OF SPEECHSENTIO**

1. **Integration**

The final step involves integrating the stutter classification and speech emotion recognition modules into a unified system. This integration ensures seamless interaction between the two modules and enables the system to analyze speech samples for both stuttering and emotional content simultaneously. Compatibility between the modules is ensured to facilitate smooth data flow and processing.

1. **Evaluation**

Following integration, the overall performance of the system is evaluated. This includes assessing its accuracy, efficiency, and effectiveness in real-world scenarios. Performance metrics specific to each module, as well as metrics reflecting the system's overall functionality, are considered during evaluation. Any shortcomings or areas for improvement identified during evaluation are addressed in the subsequent optimization phase.

1. **Optimization and deployment**

In the optimization phase, the system is fine-tuned to improve its performance. This may involve adjusting model parameters, optimizing algorithms, or refining data preprocessing techniques. Once optimized, the system is deployed in a suitable environment for real-world usage. Factors such as scalability, reliability, and maintainability are taken into account to ensure the deployed solution meets the needs of its intended users.

By following these steps and explanations, the implementation of stutter classification and speech emotion recognition systems can be carried out effectively, leading to a unified system capable of analyzing speech for both stuttering and emotional content.

**5.2 SOURCE CODE**

**index.html**

<!DOCTYPE html>

<html lang="en">

<head>

    <meta charset="UTF-8">

    <meta name="viewport" content="width=device-width, initial-scale=1.0">

    <title>Speechsentio</title>

    <link rel="stylesheet" href="{{ url\_for('static', filename='style.css') }}">

    <style>

        /\* About Us Box Styles \*/

        .about-us {

            background-color: #F2D4CF; /\* White background \*/

            color: #000000; /\* Black text color \*/

            padding: 20px;

            border-radius: 8px;

            box-shadow: 0 0 10px rgba(0, 0, 0, 0.1);

            max-width: 600px; /\* Adjust as needed \*/

            width: 90%; /\* Adjust as needed \*/

            margin: 60px auto; /\* Center horizontally with some margin \*/

             /\* Align text to the left \*/

            text-align: justify;

        }

        .about-us h2 {

            color: #ff6f61; /\* Custom color for the title \*/

            font-size: 24px; /\* Adjust as needed \*/

            margin-bottom: 10px;

        }

        /\* Dynamic Effects \*/

        .about-us {

            opacity: 0;

            transform: translateY(20px);

            transition: opacity 0.5s ease, transform 0.5s ease;

        }

        .about-us.show {

            opacity: 1;

            transform: translateY(0);

        }

    </style>

</head>

<body>

    <h1>

        &#127908; Speechsentio &#127908;

    </h1>

    <form action="/upload" method="post" enctype="multipart/form-data">

        <label for="audioFile">Upload Audio File:</label>

        <input type="file" name="audioFile" id="audioFile" accept=".wav" required>

        <br>

        <button type="submit">Submit</button>

    </form>

    {% if message %}

        <p>{{ message }}</p>

    {% endif %}

    <!-- About Us Box -->

    <div class="about-us">

        <h2>About Us</h2>

        <p>Does stuttering hold you back from expressing yourself? SpeechSentio can help! Stuttering can arise from neurological factors, but emotions like anxiety can also make it worse. SpeechSentio empowers you by providing techniques to manage emotions and build confidence. Calmer speech often leads to smoother speech. Through practice exercises, you'll develop the tools you need to communicate clearly and confidently, leaving a strong impression every time..</p>

    </div>

    <!-- JavaScript for Dynamic Effects -->

    <script>

        window.addEventListener('DOMContentLoaded', function() {

            var aboutUsBox = document.querySelector('.about-us');

            if (aboutUsBox) {

                aboutUsBox.classList.add('show');

            }

        });

    </script>

</body>

</html>

**style.css**

/\* General styles \*/

body {

    font-family: Georgia, 'Times New Roman', Times, serif;

    margin: 0;

    padding: 0;

    background-image: url("pic2.jpeg");

    background-size: cover; /\* Cover the entire background \*/

    background-repeat: no-repeat; /\* Prevent the background image from repeating \*/

    background-position: center; /\* Center the background image \*/

}

/\* Styles for h1 \*/

h1 {

    text-align: center;

    font-family: 'Pacifico', cursive; /\* Use a decorative font for a more unique look \*/

    font-size: 70px;

    margin-top: 20px; /\* Adjust as needed \*/

    color: #fff; /\* Text color \*/

    position: relative; /\* Positioning for dynamic effects \*/

}

/\* Create a dynamic gradient background \*/

h1::before {

    content: 'Speechsentio';

    position: absolute;

    top: 0;

    left: 0;

    width: 100%;

    height: 100%;

    background: linear-gradient(to right, #ff8c69, #ff6f61); /\* Gradient background \*/

    -webkit-background-clip: text; /\* Clip text to gradient (for webkit browsers) \*/

    background-clip: text; /\* Clip text to gradient (standard) \*/

    z-index: -1; /\* Ensure the gradient is behind the text \*/

}

/\* Add a subtle animation \*/

h1::after {

    content: '';

    position: absolute;

    top: 0;

    left: 0;

    width: 100%;

    height: 100%;

    background: linear-gradient(to right, transparent, #fff, transparent); /\* Gradient overlay \*/

    z-index: -2; /\* Ensure the gradient is behind the text \*/

    animation: shine 2s infinite linear; /\* Animation effect \*/

}

@keyframes shine {

    0% {

        transform: translateX(-100%);

    }

    100% {

        transform: translateX(100%);

    }

}

/\* Add a text shadow for depth \*/

h1 {

    text-shadow: 8px 8px 16px rgba(0, 0, 0, 0.5); /\* Increased text shadow for depth \*/

}

/\* Add hover effect for interactivity \*/

h1:hover {

    color: #ff6f61; /\* Change text color on hover \*/

}

.container {

    background-color: #ff6f61; /\* Attractive background color \*/

    background-image: url("pic2.jpeg"); /\* Image URL relative to the HTML file \*/

    background-size: cover;

    background-position: center;

    display: flex;

    flex-direction: column;

    align-items: center;

    justify-content: center;

    height: 100vh;

    margin: auto; /\* Center the container horizontally \*/

    text-align: center; /\* Center text elements horizontally within the container \*/

    box-shadow: 0 0 20px rgba(0, 0, 0, 0.5); /\* Increased box shadow \*/

}

form {

    background-color: #ffffff; /\* Solid white background \*/

    color: #ffffff; /\* White text color \*/

    padding: 20px;

    border-radius: 8px;

    box-shadow: 0 0 10px rgba(0, 0, 0, 0.1);

    max-width: 400px; /\* Adjust as needed \*/

    width: 100%; /\* Ensure the form takes full width of the container \*/

    margin: 0 auto; /\* Center the form horizontally \*/

    margin-top: 20px; /\* Adjust as needed \*/

}

form label,

form input[type="file"] {

    color: #000000; /\* Black text color \*/

}

label {

    display: block;

    margin-bottom: 10px;

}

input[type="file"] {

    width: 94.5%;

    padding: 10px;

    margin-bottom: 10px;

    border: 1px solid #ccc;

    border-radius: 4px;

}

button {

    background-color: #4CAF50;

    color: white;

    padding: 10px 20px;

    border: none;

    border-radius: 4px;

    cursor: pointer;

    width: 100%;

}

button:hover {

    background-color: #45a049;

}

/\* Message styles \*/

p {

    text-align: center;

    margin-top: 20px;

    color: #fff; /\* Text color \*/

}

.result-message {

    margin-top: 20px;

    padding: 20px;

    background-color: rgba(255, 255, 255, 0.8);

    border-radius: 8px;

    box-shadow: 0 0 10px rgba(0, 0, 0, 0.1);

    max-width: 600px;

    width: 100%;

    margin: 0 auto;

    text-align: left;

    color: #333; /\* Adjust text color \*/

}

/\* About Us Box Styles \*/

.about-us {

    background-color: #ffffff; /\* White background \*/

    color: #000000; /\* Black text color \*/

    padding: 20px;

    border-radius: 8px;

    box-shadow: 0 0 10px rgba(0, 0, 0, 0.1);

    max-width: 600px; /\* Adjust as needed \*/

    width: 90%; /\* Adjust as needed \*/

    margin: 20px auto; /\* Center horizontally with some margin \*/

    text-align: left; /\* Align text to the left \*/

    opacity: 0; /\* Initially hidden \*/

    transform: translateY(20px); /\* Initially moved down \*/

    transition: opacity 0.5s ease, transform 0.5s ease; /\* Smooth transition \*/

}

.about-us h2 {

    color: #ff6f61; /\* Custom color for the title \*/

    font-size: 24px; /\* Adjust as needed \*/

    margin-bottom: 10px;

}

/\* Show the About Us box when it becomes visible \*/

.about-us.show {

    opacity: 1;

    transform: translateY(0);

}

/\* About Us Box Styles \*/

.about-us {

    background-color:#F2D4CF; /\* White background \*/

    color: #000000; /\* Black text color \*/

    padding: 20px;

    border-radius: 8px;

    box-shadow: 0 0 10px rgba(0, 0, 0, 0.1);

    max-width: 600px; /\* Adjust as needed \*/

    width: 90%; /\* Adjust as needed \*/

    margin: 60px auto; /\* Center horizontally with some margin \*/

    text-align: left; /\* Align text to the left \*/

    opacity: 0; /\* Initially hidden \*/

    transform: translateY(20px); /\* Initially moved down \*/

    transition: opacity 0.5s ease, transform 0.5s ease; /\* Smooth transition \*/

}

.about-us h2 {

    color: #ff6f61; /\* Custom color for the title \*/

    font-size: 24px; /\* Adjust as needed \*/

    margin-bottom: 10px;

}

/\* Set text color for paragraphs inside the About Us box \*/

.about-us p {

    color: #000000;

    text-align: justify;/\* Black text color \*/

    font-family: 'Roboto', sans-serif;

}

/\* Show the About Us box when it becomes visible \*/

.about-us.show {

    opacity: 1;

    transform: translateY(0);

}

**Result.html**

<!DOCTYPE html>

<html lang="en">

<head>

    <meta charset="UTF-8">

    <meta name="viewport" content="width=device-width, initial-scale=1.0">

    <title>Result</title>

    <style>

        body {

            background-size: cover;

            background-repeat: no-repeat;

            font-family: Arial, sans-serif;

        }

        .container {

            margin: auto;

            width: 50%;

            border: 1px solid #ccc;

            padding: 20px;

            background-color: #ffffff; /\* Solid white background \*/

        }

        .container h2 {

            text-align: center;

            color: #333;

            font-family: Cambria, Cochin, Georgia, Times, 'Times New Roman', serif;

            font-size: 30px;

            margin: 0;

            padding-bottom: 10px;

        }

        .container .content {

            font-size: 16px;

        }

        .container .sub-title {

            font-size: 24px;

            text-decoration: underline;

            margin-top: 20px;

            margin-bottom: 10px;

        }

        .green-container {

            background-color: #9b59b6;

            padding: 10px;

            margin-bottom: 20px;

            text-align: center;

        }

    </style>

</head>

<body style="background-image: url('{{ url\_for('static', filename='pic2.jpeg') }}');">

    <div class="container">

        <div class="green-container">

            <h2>Audio Analysis and Recommendation</h2>

        </div>

        <h3 class="sub-title">Audio Analysis</h3>

        <div class="content">{{ message | safe }}</div>

    </div>

</body>

</html>

**app.py**

from flask import Flask, render\_template, request

from tensorflow.keras.models import load\_model

import numpy as np

import librosa

import soundfile as sf

import os

from markupsafe import Markup

import pickle

import random

import time

app = Flask(\_\_name\_\_)

# Load your stutter classification models

word\_repetition\_model = load\_model(r"C:\Users\BEHARA AMULYA\Downloads\Major project\stutter-classification-main\prolongation\_model.h5")

sound\_repetition\_model = load\_model(r"C:\Users\BEHARA AMULYA\Downloads\Major project\stutter-classification-main\sound\_repetition\_model.h5")

prolongation\_model = load\_model(r"C:\Users\BEHARA AMULYA\Downloads\Major project\stutter-classification-main\prolongation\_model.h5")

# Load the emotion recognition model

emotion\_model = pickle.load(open(r"C:\Users\BEHARA AMULYA\Downloads\Major project\emotion\_model.pkl", "rb"))

# Emotion mapping

emotion\_mapping = {'0': 'Neutral', '1': 'Calm', '2': 'Happy', '3': 'Sad', '4': 'Angry', '5': 'Fearful', '6': 'Disgust', '7': 'Surprised'}

# Function to extract features from audio file

def extract\_features(file\_path):

    X, sample\_rate = librosa.load(file\_path, res\_type='kaiser\_fast')

    mfccs = np.mean(librosa.feature.mfcc(y=X, sr=sample\_rate, n\_mfcc=13).T, axis=0)

    return mfccs.reshape(1, -1)

def extract\_feature(file\_name, mfcc=True, chroma=True, mel=True):

    with sf.SoundFile(file\_name) as sound\_file:

        X = sound\_file.read(dtype="float32")

        sample\_rate = sound\_file.samplerate

        if chroma:

            stft = np.abs(librosa.stft(X))

        result = np.array([])

        if mfcc:

            mfccs = np.mean(librosa.feature.mfcc(y=X, sr=sample\_rate, n\_mfcc=40).T, axis=0)

            result = np.hstack((result, mfccs))

        if chroma:

            chroma = np.mean(librosa.feature.chroma\_stft(S=stft, sr=sample\_rate).T, axis=0)

            result = np.hstack((result, chroma))

        if mel:

            mel = np.mean(librosa.feature.melspectrogram(y=X, sr=sample\_rate).T, axis=0)

            result = np.hstack((result, mel))

        pad\_size = 180 - len(result)

        if pad\_size > 0:

            result = np.pad(result, (0, pad\_size), 'constant')

    return result

# Function to classify stutter type

def classify\_stutter(file\_path, model):

    features = extract\_features(file\_path)

    prediction = model.predict(features)

    return prediction[0]

# Function to predict emotion

def predict\_emotion(file\_path):

    feature = extract\_feature(file\_path)

    feature = feature.reshape(1, -1)

    emotion\_prediction = emotion\_model.predict(feature)

    return emotion\_prediction[0]

# Function to generate a random paragraph

def generate\_random\_paragraph():

    sentences = [

        "The quick brown fox jumps over the lazy dog.",

        "A journey of a thousand miles begins with a single step.",

        "All that glitters is not gold.",

        "In the midst of winter, I found there was, within me, an invincible summer.",

        "To be yourself in a world that is constantly trying to make you something else is the greatest accomplishment.",

        "Life is what happens when you're busy making other plans.",

        "The only limit to our realization of tomorrow will be our doubts of today.",

        "Success is not final, failure is not fatal: It is the courage to continue that counts.",

        "The greatest glory in living lies not in never falling, but in rising every time we fall.",

        "The purpose of our lives is to be happy."

    ]

    num\_sentences = random.randint(3, 6)

    paragraph = " ".join(random.sample(sentences, num\_sentences))

    return paragraph

def get\_random\_tongue\_twister():

    tongue\_twisters = [

        "How much wood would a woodchuck chuck\nIf a woodchuck could chuck wood?\nHe would chuck as much wood as a woodchuck would\nIf a woodchuck could chuck wood.",

        "I saw Susie sitting in a shoeshine shop.\nWhere she sits she shines, and where she shines she sits.",

        "She sells seashells by the seashore,\nThe shells she sells are surely seashells.\nSo if she sells shells on the seashore,\nI'm sure she sells seashore shells.",

        "Betty Botter bought some butter,\nBut she said the butter's bitter.\nIf I put it in my batter,\nIt will make my batter bitter.\nSo she bought some better butter,\nBetter than the bitter butter.\nAnd she put it in her batter,\nAnd her batter was not bitter.\nSo 'twas better Betty Botter\nBought some better butter.",

        "A proper copper coffee pot.\nI'm not a pheasant plucker,\nI'm a pheasant plucker's son.\nI'm only plucking pheasants\nTill the pheasant plucker comes.",

        "Six slippery snails slid silently seaward.\nSilly Sally swiftly shooed seven silly sheep.\nThe seven silly sheep Silly Sally shooed\nShilly-shallied south.\nThese sheep shouldn't sleep in a shack;\nSheep should sleep in a shed.",

        "I saw Susie sitting in a shoeshine shop.\nWhere she sits she shines, and where she shines she sits.",

        "How can a clam cram in a clean cream can?\nIf you must cross a course, cross cow across a crowded cow crossing,\nCross the cross coarse cow across the crowded cow crossing carefully.",

        "Fuzzy Wuzzy was a bear.\nFuzzy Wuzzy had no hair.\nFuzzy Wuzzy wasn't very fuzzy, was he?",

        "Near an ear, a nearer ear, a nearly eerie ear.\nA proper cup of coffee in a copper coffee cup."

    ]

    return random.choice(tongue\_twisters)

def interpret\_predictions(word\_repetition\_output, sound\_repetition\_output, prolongation\_output, emotion\_output):

    audio\_analysis = "Audio Analysis:<br>"

    recommendations = "Recommendation:<br>"

    audio\_analysis += "Word Repetition Prediction: "

    if word\_repetition\_output > 0.05:

        audio\_analysis += "Word Repetition found"

        recommendations += "You are repeating the same word again and again. Practice slow and deliberate speech. Take a pause between words to allow for smoother communication.<br>"

    else:

        audio\_analysis += "No Word Repetition"

    audio\_analysis += "<br><br>"

    audio\_analysis += "Sound Repetition Prediction: "

    if sound\_repetition\_output > 0.01:

        audio\_analysis += "Sound Repetition found"

        recommendations += "Sound Repetition is found in your voice. Work on relaxation techniques to reduce tension. Focus on breathing exercises to promote a more natural speech flow.<br>"

    else:

        audio\_analysis += "No Sound Repetition"

    audio\_analysis += "<br><br>"

    audio\_analysis += "Prolongation Prediction: "

    if prolongation\_output > 0.05:

        audio\_analysis += "Prolongation found"

        recommendations += "And Prolongation is evident in your voice. Practice controlled breathing to ease into speech sounds. Gradually increase the speed of speech while maintaining control.<br>"

    else:

        audio\_analysis += "No Prolongation"

    audio\_analysis += "<br><br>"

    audio\_analysis += f"Emotion Prediction: {emotion\_output}"

    audio\_analysis += "<br><br>"

    # Check if any stutter type is detected, if not, print random paragraphs and tongue twisters

    if word\_repetition\_output > 0.05 or sound\_repetition\_output > 0.05 or prolongation\_output > 0.05:

        recommendations += f"<br>Practice tongue twisters like these regularly to improve your speech:<br>"

        for i in range(3):

            recommendations += f"{i + 1}. {get\_random\_tongue\_twister()}<br>"

        recommendations += "<br>Improve your fluency by reading these sentences slowly and clearly:<br>"

        for i in range(3):

            recommendations += f"{i + 1}. {generate\_random\_paragraph()}<br>"

    return Markup(audio\_analysis + recommendations)

@app.route('/')

def home():

    return render\_template('index.html')

@app.route('/upload', methods=['POST'])

def upload():

    if 'audioFile' not in request.files:

        return render\_template('index.html', message='No file part')

    file = request.files['audioFile']

    if file.filename == '':

        return render\_template('index.html', message='No selected file')

    if file:

        file\_path = 'uploads/' + file.filename

        file.save(file\_path)

        word\_repetition\_output = classify\_stutter(file\_path, word\_repetition\_model)

        sound\_repetition\_output = classify\_stutter(file\_path, sound\_repetition\_model)

        prolongation\_output = classify\_stutter(file\_path, prolongation\_model)

        emotion\_output = predict\_emotion(file\_path)

        result\_message = interpret\_predictions(word\_repetition\_output, sound\_repetition\_output, prolongation\_output, emotion\_output)

        return render\_template('result.html', message=result\_message, prediction=emotion\_output)

@app.route('/result')

def result():

    return render\_template('result.html')

if \_\_name\_\_ == '\_\_main\_\_':

    app.run(debug=True)

**CHAPTER 6**

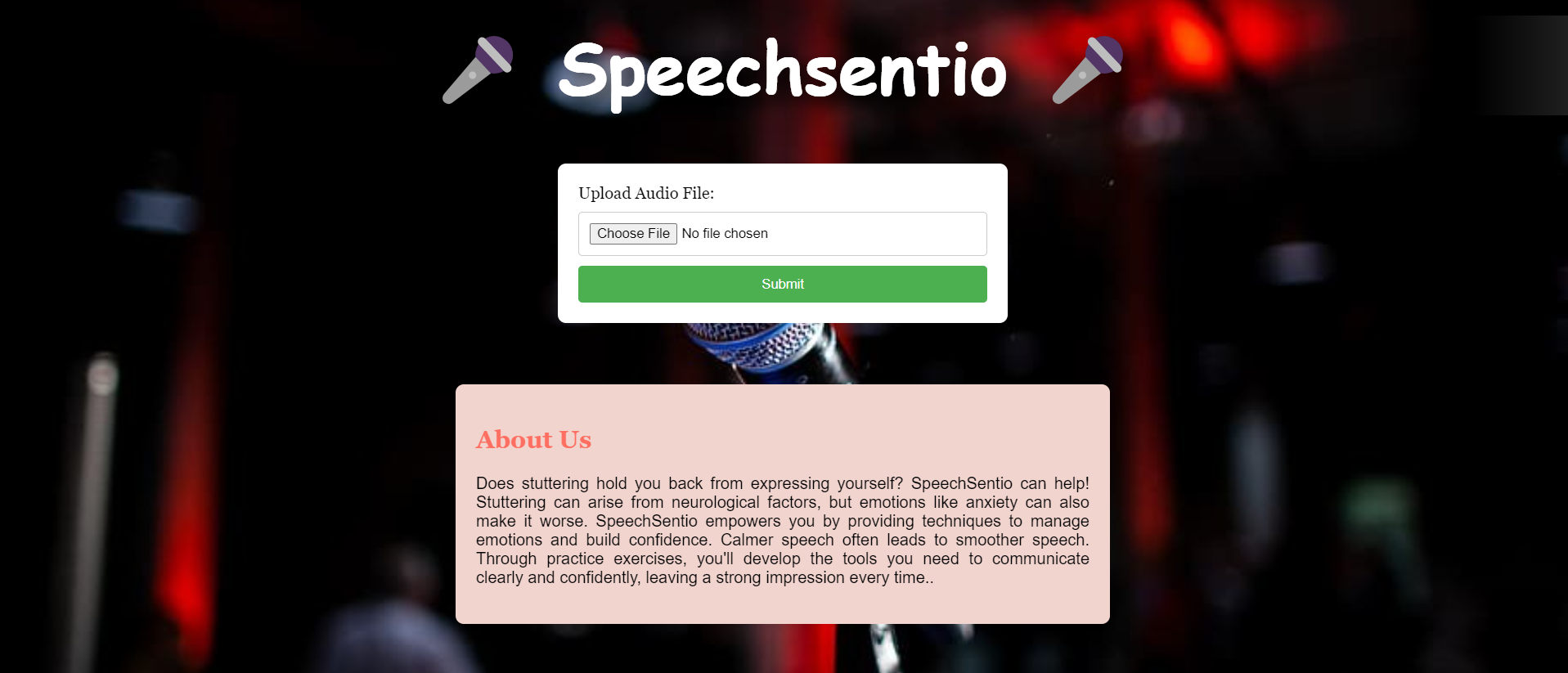
**RESULTS**

The results showcase promising accuracies in stutter detection and emotion recognition, with algorithms like Decision Tree achieving 79.5% accuracy and K-Nearest Neighbors achieving 85% accuracy. Additionally, the Multi-Layer Perceptron (MLP) model demonstrates notable performance, achieving 78% accuracy, highlighting the effectiveness of neural networks in speech therapy applications. These results underscore the potential of advanced machine learning techniques to enhance the diagnosis and treatment of speech disorders, paving the way for more tailored and effective therapeutic interventions.

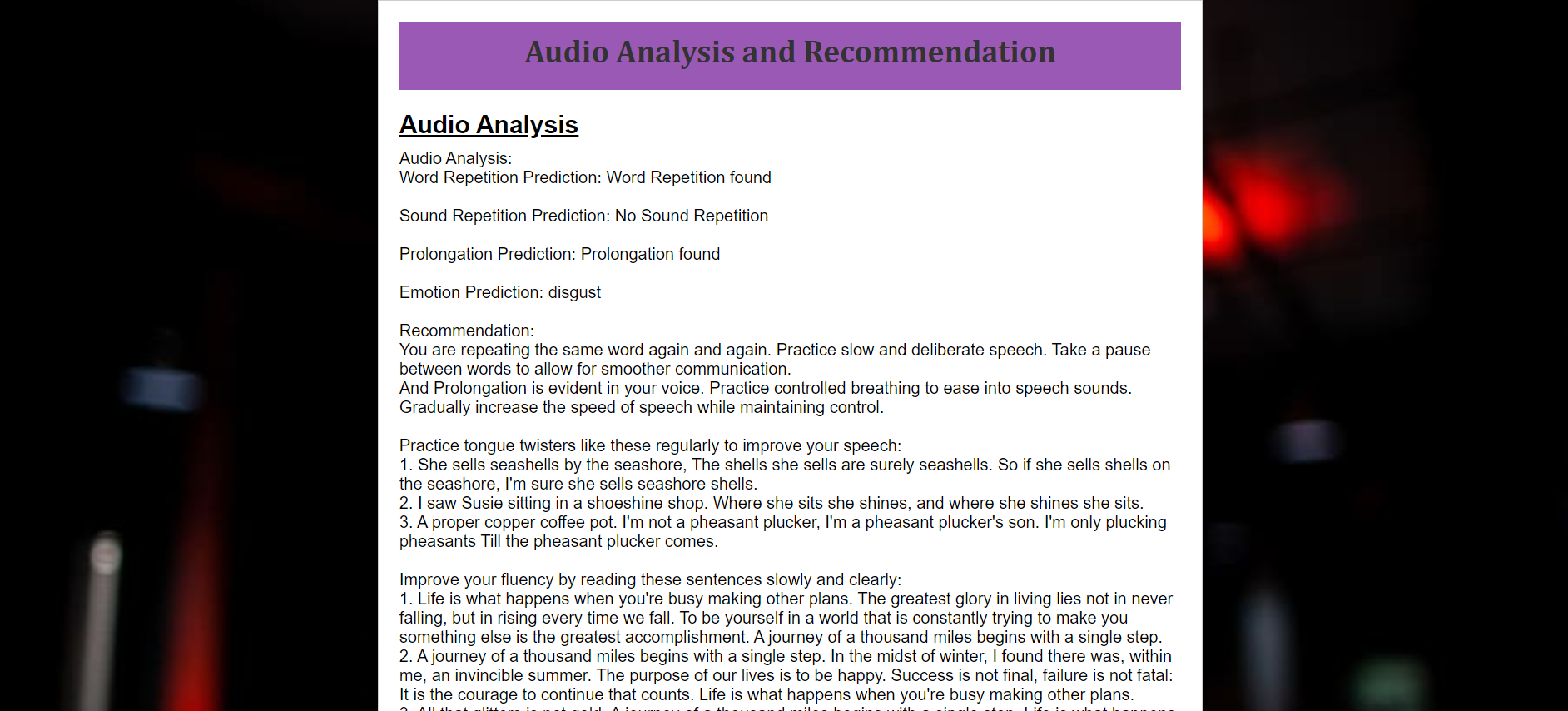
Top of Form

SpeechSentio utilizes a dual-branch deep learning architecture to analyze speech in real-time. One branch detects stutters using algorithms like K-Nearest Neighbors, Decision tree while the other recognizes emotions through features extracted from speech patterns.

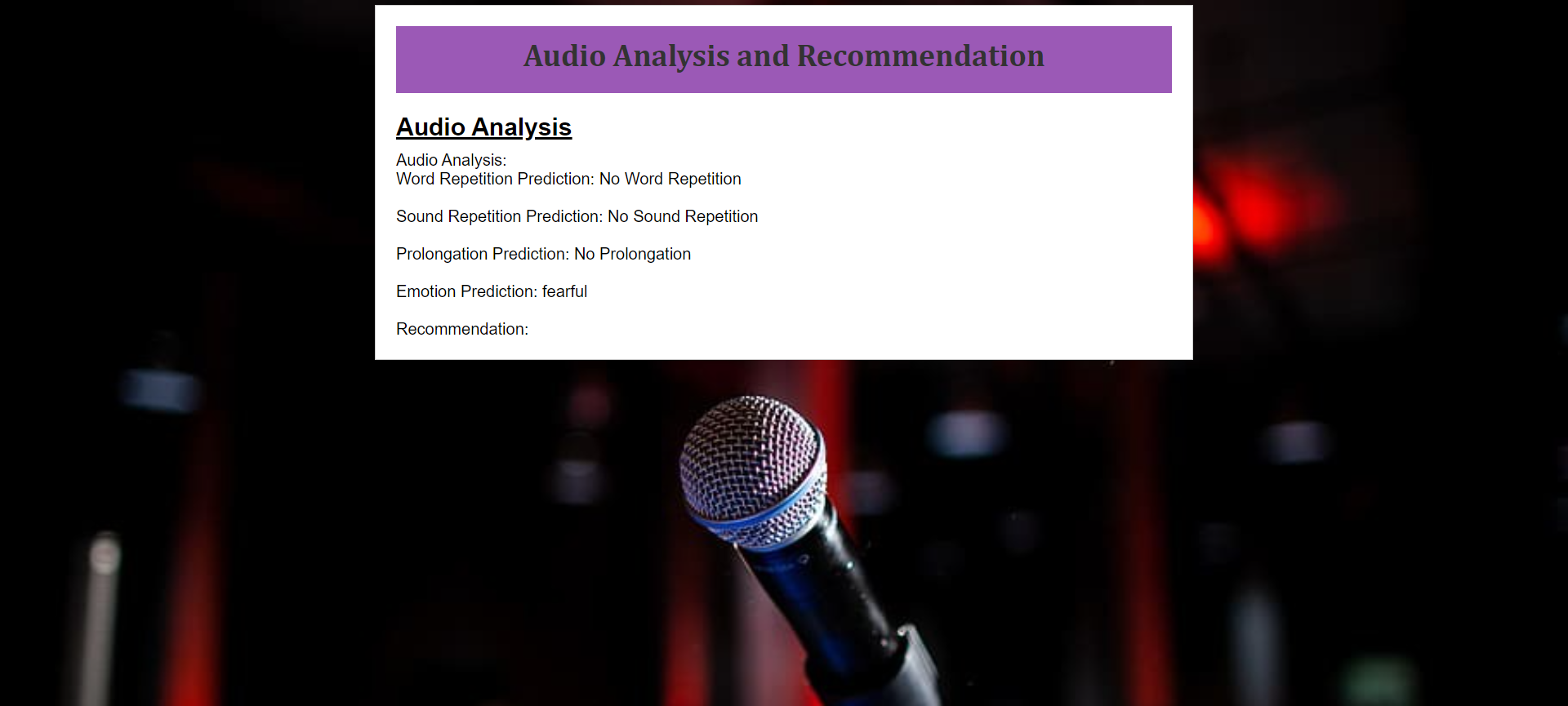
The obtained results after the development of the SpeechSentio application are as follows.



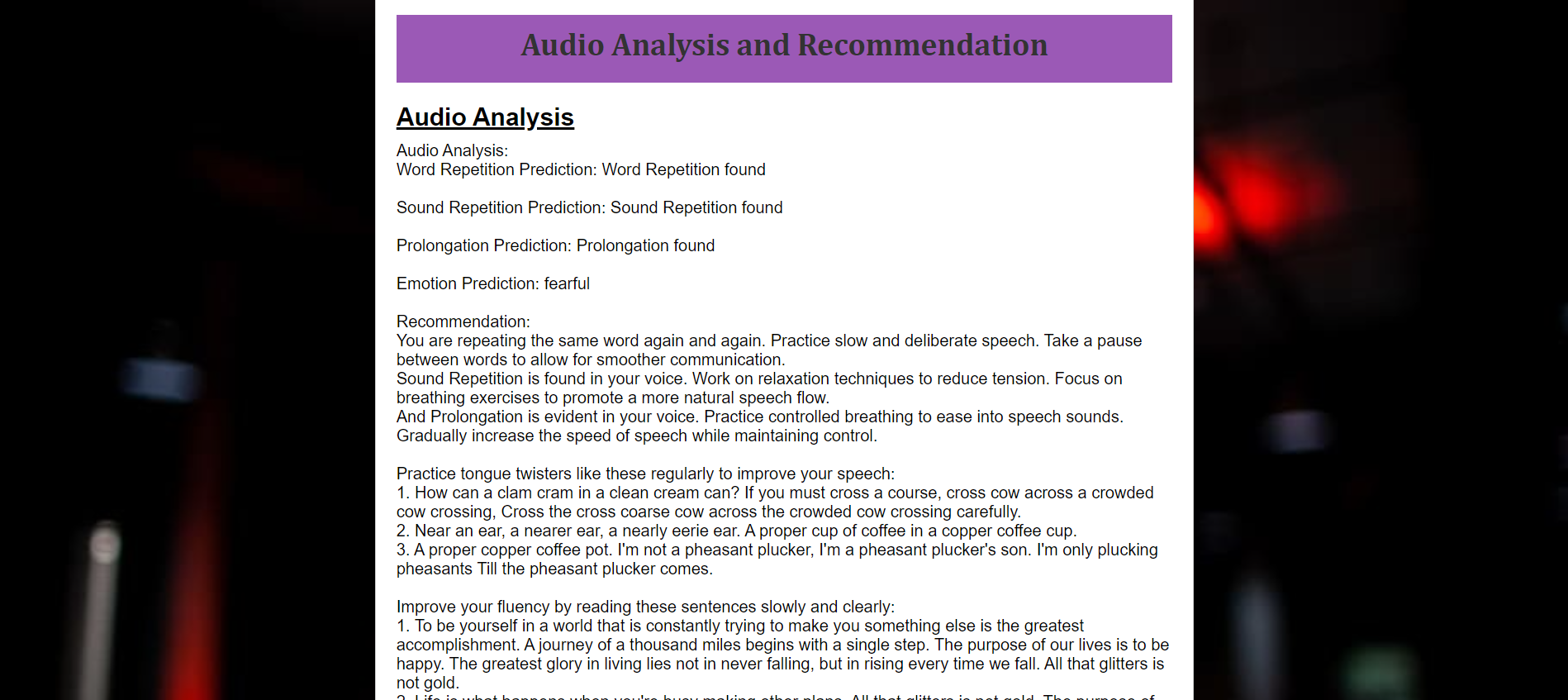
**Figure 13: SpeechSentio speech pattern feedback via user-uploaded audio**



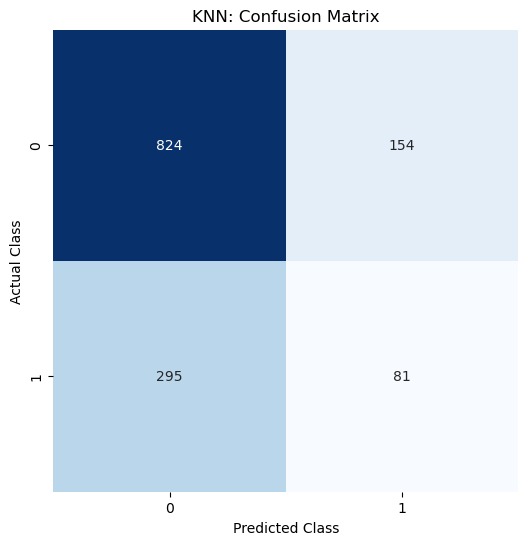
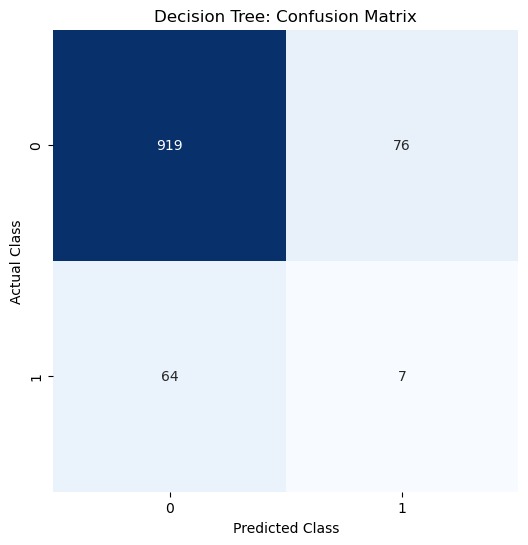
**Figure 14: SpeechSentio Audio Analysis Report for “WomenWhoStutter\_97\_82.wav”**



**Figure 15: SpeechSentio Audio Analysis Report for “WomenWhoStutter\_99\_163.wav”**

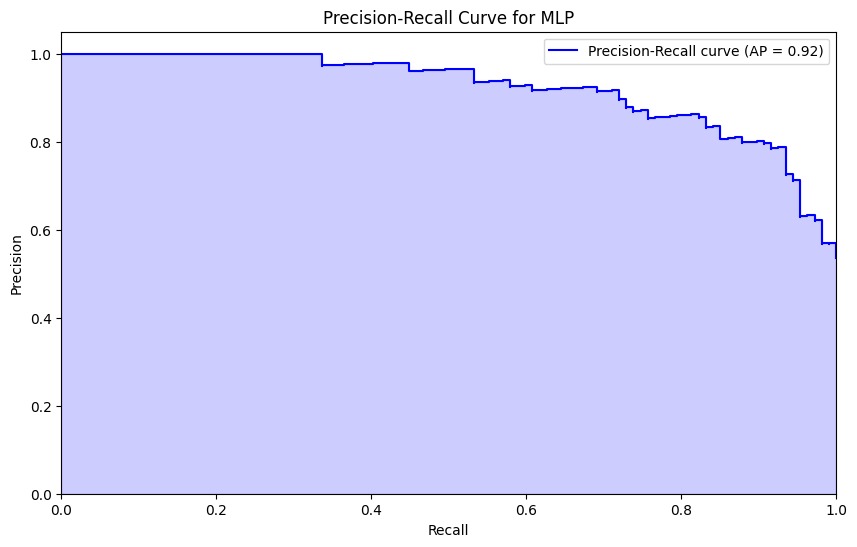


**Figure 16: SpeechSentio Audio Analysis Report for “03-01-07-01-01-01-04.wav”**





**Figure 17: Confusion matrix for Decision Tree, KNN and MLP**



**Figure 18: Precision-Recall Curve for MLP**

| S. No | ML algorithms | | |
| --- | --- | --- | --- |
| Algorithm | Accuracy using training dataset | Accuracy using testing dataset |
| 1. | Decision Tree | 80.375% | 79.5% |
| 2. | K- Nearest Neighbors | 82% | 85% |
| 3. | Multi-Layer Perceptron | 75.35% | 78% |

**Table 6.1 Accuracies obtained on ML algorithms**

**CHAPTER 7**

**CONCLUSION**

SpeechSentio presents a novel approach to stutter detection and speech therapy by integrating machine learning, signal processing, and emotion analysis. Through the incorporation of emotional cues, our framework enhances the accuracy of stutter identification and offers valuable insights into the emotional aspects of speech disorders. By leveraging features extracted from audio signals and employing decision trees and KNN for stutter classification, SpeechSentio demonstrates promising results in identifying different types of stuttering. Additionally, our use of MLP for emotion recognition further enriches the framework's capabilities. With its potential to revolutionize human-computer interactions and mental health assessments, SpeechSentio signifies a significant advancement in the field of speech therapy and holds promise for improving therapeutic interventions and personalized treatment approaches.

**FUTURE ENHANCEMENTS AND DISCUSSIONS**

Future enhancements could focus on refining the emotion recognition aspect of SpeechSentio by exploring more advanced neural network architectures and incorporating larger and more diverse emotional datasets. Additionally, further research could investigate the integration of real-time feedback mechanisms to provide immediate support and guidance during speech therapy sessions. Moreover, expanding the framework to encompass other speech disorders beyond stuttering could broaden its applicability and impact in the field of speech therapy. Collaborations with speech therapists and psychologists could provide valuable insights for fine-tuning the framework and tailoring it to meet the specific needs of individuals with speech disorders.