HIGH LEVEL DESIGN DOCUMENT

FeelSpeak: Generating Emotional Speech with Deep Learning

UE20CS390A – Capstone Project Phase – 1

Submitted by:

M H SOHAN	**PES1UG20CS235**
RAHUL ROSHAN G	**PES1UG20CS320**
ROHIT ROSHAN	**PES1UG20CS355**
S M SUTHARSAN RAJ	**PES1UG20CS362**

Under the guidance of

Prof. V R BADRI PRASAD
Associate Professor
PES University

January - May 2023

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
FACULTY OF ENGINEERING
PES UNIVERSITY
(Established under Karnataka Act No. 16 of 2013)
100ft Ring Road, Bengaluru – 560 085, Karnataka, India

HIGH LEVEL DESIGN DOCUMENT

TABLE OF CONTENTS

HIGH LEVEL DESIGN DOCUMENT

1. Introduction

A text editor with audio speech is a software tool that allows users to create and edit text documents using voice commands. It uses speech recognition technology to convert spoken words into written text, allowing users to dictate and edit text without the need for a keyboard or mouse.

Currently these text editors have robotic voices when it reads out the text in the editor. Due to this , the tool has very little practical application in the real world. Our idea is to bring out the tool to the text editor which could read out the text with emotion .The emotion includes happy,sad,angry, neutral.sarcasm.

This would bring out many practical application:

Used in classroom,Office meeting, read out to people with visual impairment

High level design

- The input data consist of dataset from various sources like social media, poetry, paragraph .Preprocessing is done on the text dataset like removing stop words, explicit words,punctuation.
- The labeled text data being provided as input to the emotion detection training model. The model analyzes the text and identifies the corresponding emotion. This output is then passed on to the speech emotion detection model.
- The speech emotion detection model receives speech data that has been labeled with an emotion. The model analyzes the speech and identifies the corresponding emotion. This output is also passed on to the speech emotion model.
- The labeled text data is then passed through the speech emotion model. The model adjusts the speech to reflect the appropriate emotion based on the input text and the corresponding emotion identified by the emotion detection training model.
- The resulting output is speech that accurately reflects the emotional content of the input text. This scenario can be repeated for multiple input texts, allowing the system to generate emotional speech for a variety of different inputs.

2. Current System [if applicable]

Deep learning systems and projects typically utilize emotion detection models that recognize emotions such as anger, sadness, happiness, and neutral, but very few incorporate sarcasm as an emotion. Our proposed idea involves incorporating sarcasm

as an emotion in an emotion detection model, which would then be integrated into a text editor that can read out the emotion with a sarcastic tone, in addition to anger, sadness, happiness, and neutral. Our goal with this project is to enhance the accuracy of models that currently lack sarcasm as an emotion category.

3. Design Considerations

3.1. Design Goals

- Our proposed idea involves incorporating sarcasm as an emotion in an emotion detection model, which would then be integrated into a text editor.
- This feature can read out the emotion with a sarcastic tone, in addition to anger, sadness, happiness, and neutral.
- Our goal with this project is to enhance the accuracy of models that currently lack sarcasm as an emotion category.

3.2. Architecture Choices

Alternative choices considered for the proposed system include using pre-trained emotion detection models, using only speech data without incorporating text data, and using rule-based approaches to detect emotions in text and speech.

Using pre-trained emotion detection models would require less training time but may not accurately reflect the nuances of sarcasm and other less common emotions. Using only speech data without incorporating text data would limit the system's ability to accurately capture the intended emotional content of the text. Using rule-based approaches would require significant domain expertise and may not be able to capture the complexity and variability of emotional expression.

The proposed system of incorporating both text and speech data and training a model specifically for detecting sarcasm, in addition to other emotions, is the most appropriate choice because it takes into account the limitations of the other alternatives and provides a more comprehensive and accurate approach to detecting emotions. By training a model specifically for sarcasm, the system can more accurately capture the nuances of this emotion, which may be missed by other models.

Pros

The proposed system includes the ability to accurately capture a wide range of emotions, including sarcasm, and the potential to improve the accuracy of existing emotion detection models.

Cons

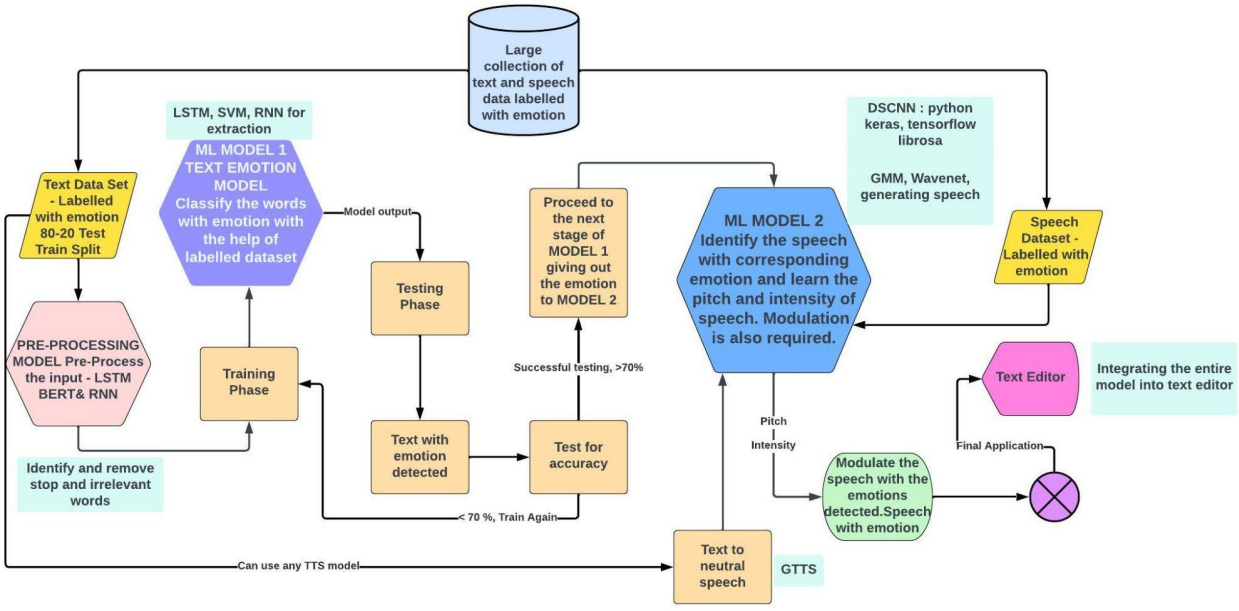
This includes the need for significant training data and the potential for errors or inaccuracies in the emotion detection and speech generation processes.

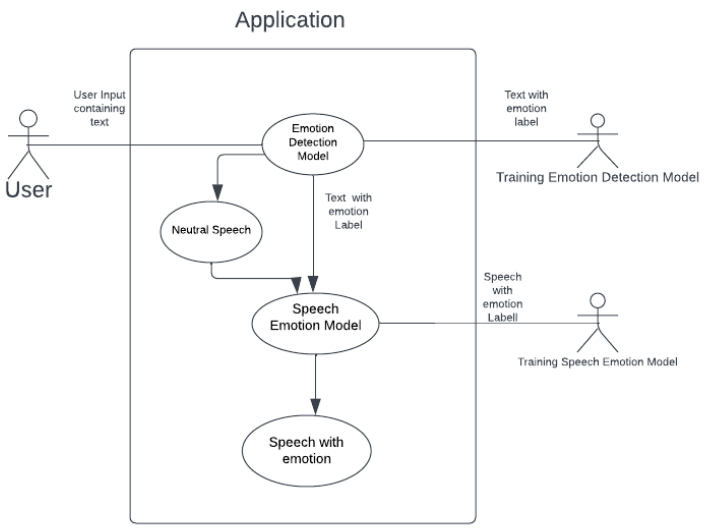
3.3. Constraints, Assumptions and Dependencies

These are the limitations, constraints that have a significant impact on the design of the system.

- Assumptions would be that the Text entered will be in English Language with a UK/US accent.
- A total of four emotions are proposed to be detected, which include happy, sad, angry, sarcastic.
- Text is assumed to be an annotated form of text, i.e, coversational format of text, role based dialogue format, etc.

4. High Level System Design





5. Design Description

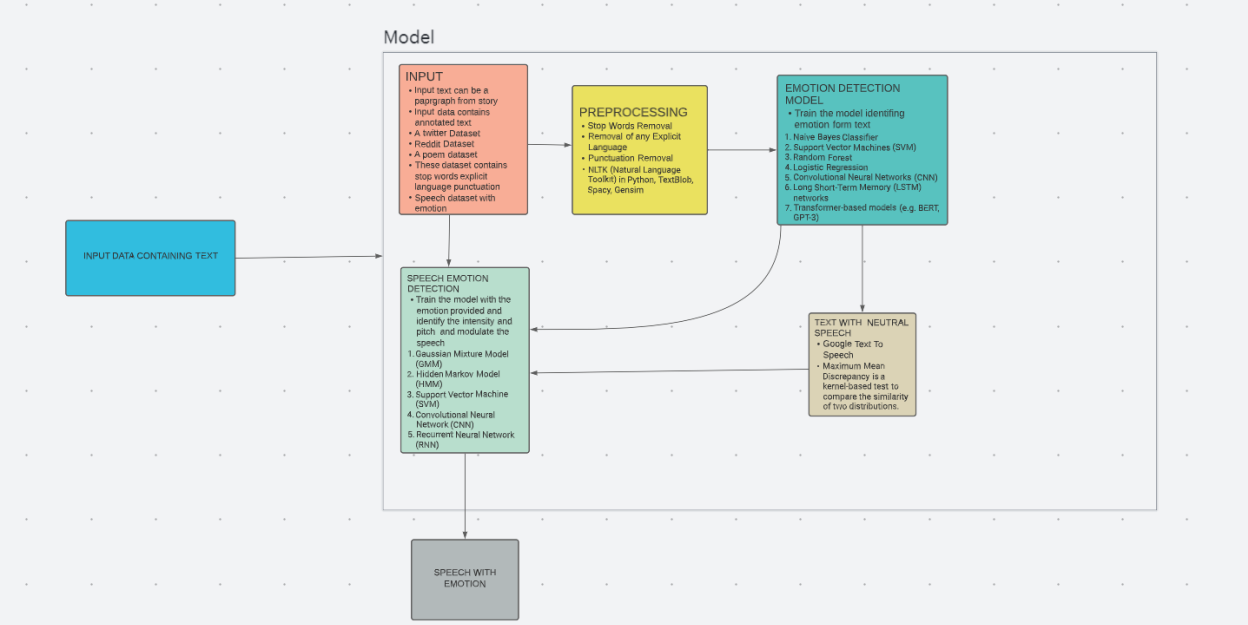
- **Text Preprocessing Module:** This module is responsible for preprocessing the input text data. It includes removing stop words, explicit words, punctuation, and other unnecessary elements from the text. The output of this module is a cleaned text that is ready for further processing.

- **Emotion Detection Training Module:** This module is responsible for training the emotion detection model. It takes the preprocessed text data as input and trains the model to recognize four different emotions: anger, sadness, happiness, and sarcasm. The output of this module is the trained emotion detection model.

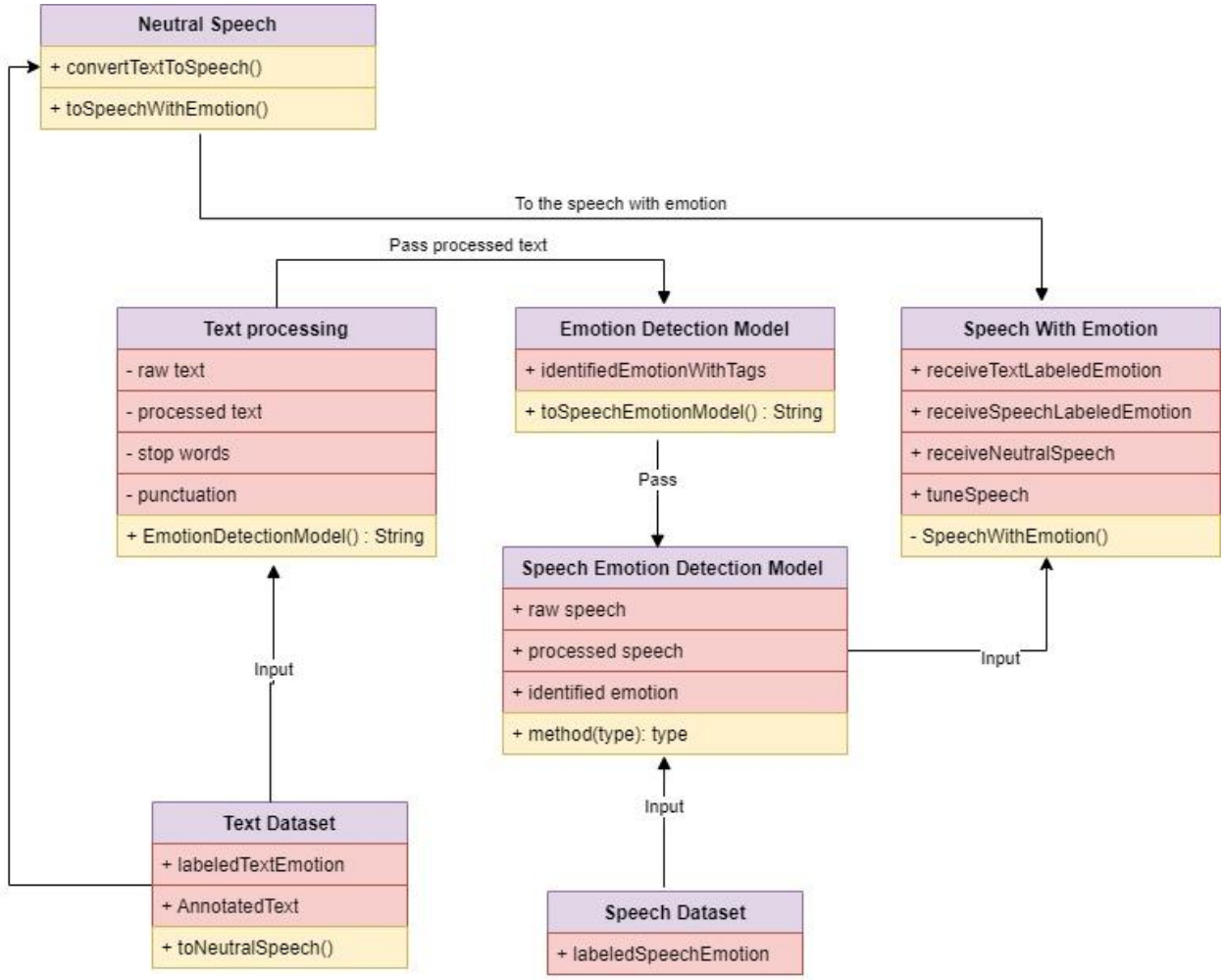
- **Speech Emotion Detection Module:** This module is responsible for detecting the emotional content of speech. It takes speech data as input and identifies the corresponding emotion. The output of this module is the emotional label of the input speech.

- **Speech Emotion Adjustment Module:** This module is responsible for adjusting the speech to reflect the appropriate emotion based on the input text and the corresponding emotion identified by the emotion detection training model. It takes

the emotional label of the input speech and the emotional label of the corresponding text as input and adjusts the speech to match the emotional content of the text.



5.1. Master Class Diagram



5.2. Reusability Considerations

For this project, we plan to use various reusable components to enhance the reusability of the project. These components include:

- **Pretrained language models**: We plan to use existing pre-trained language models such as BERT, RoBERTa, and GPT-3 to perform text analysis and emotion detection. These models have been trained on large datasets and can accurately classify text into various emotions.

- **Speech emotion detection models**: We plan to use existing speech emotion detection models to detect the emotion in the speech data. These models have been trained on large datasets and can accurately classify speech into various emotions.

- **Text preprocessing libraries:** We plan to use existing text preprocessing libraries such as NLTK and spaCy to preprocess the input text data. These libraries provide various tools for tokenization, stop word removal, stemming, and lemmatization, which can be used to clean and normalize the input text data.

- **Speech synthesis libraries**: We plan to use existing speech synthesis libraries such as Google Text-to-Speech (TTS) and Amazon Polly to generate speech from the input text data. These libraries provide high-quality synthesized speech in various languages and voices.

We also plan to build certain components within the project that can be reused, such as:

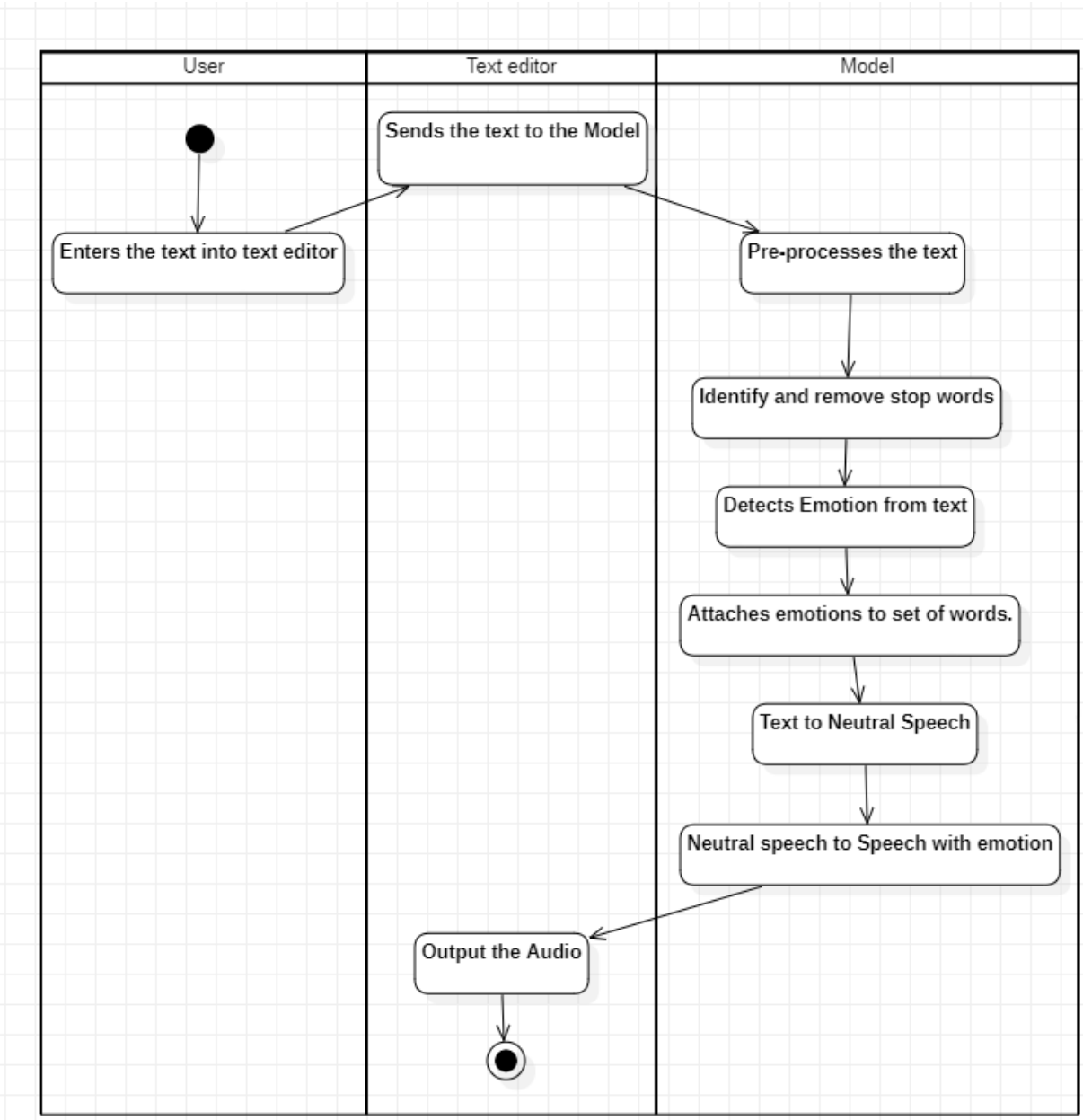
- **Custom emotion detection models:** We plan to fine-tune the pre-trained language models to improve the accuracy of emotion detection. These custom models can be reused in other projects that require emotion detection.

- **Custom speech emotion detection models:** We plan to build custom speech emotion detection models using deep learning techniques. These models can be reused in other projects that require speech emotion detection.

- **Custom text preprocessing pipelines:** We plan to build custom text preprocessing pipelines that can be reused in other projects that require text cleaning and normalization.

6. ER Diagram / Swimlane Diagram / State Diagram (include as appropriate)

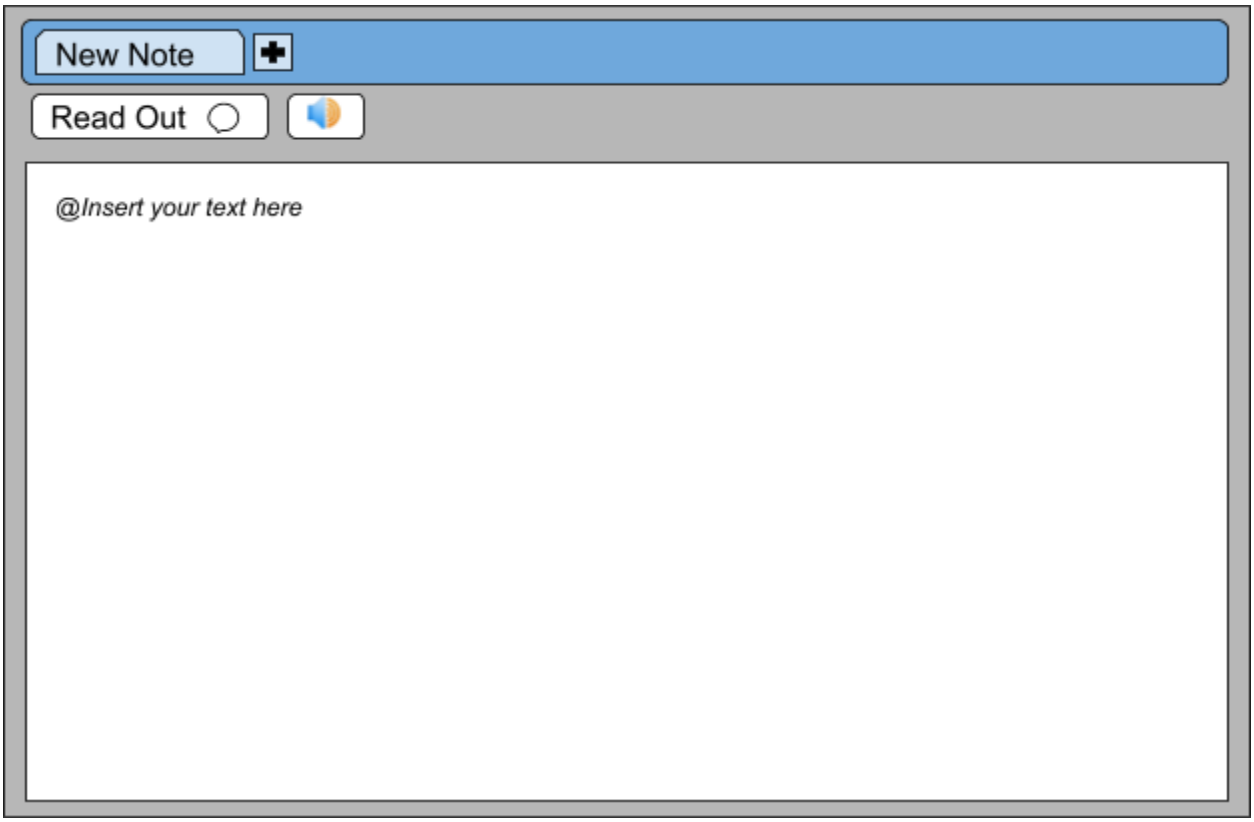
[Include the ER Diagram. The following table shall be filled for details of the entities and their data elements / attributes. Include the description of the data / function used in each module / function.]



#	Entity	Name	Definition	Type
ENTITIES				
1.	Text Dataset	Emotion	Set of text with emotion	CSV

2.	Speech Dataset	Emotional speech	Speech samples with different emotions	.WAV
#	**Attribute**	**Name**	**Definition**	**Type (size)**
DATA ELEMENTS				
1.	Emotions_1	Happy, Sad		
2.	Emotions_2	Angry, Neutral		

7. User Interface Diagrams



8. Report Layouts

Description: The report will provide insights into the emotion recognition and speech conversion process. It will contain details of the input dataset, emotion detection training model, speech emotion detection model, and speech emotion model. The report

will include a description of the system's design, algorithms used, and technologies implemented.

Selection Criteria: The report will be generated based on the input text data and the corresponding emotion identified by the emotion detection training model. It will provide a summary of the speech emotion detection model's output and the final speech output after adjusting it to reflect the appropriate emotion.

The report will contain the following tables:

- **Input Text Data:** This table will provide details of the input text dataset, including the number of records and the preprocessing techniques used.

- **Emotion Detection Training Model:** This table will provide details of the emotion detection training model's architecture, including the number of layers and neurons used.

- **Speech Emotion Detection Model:** This table will provide details of the speech emotion detection model's architecture, including the number of layers and neurons used.

- **Speech Emotion Model:** This table will provide details of the speech emotion model's architecture, including the number of layers and neurons used.

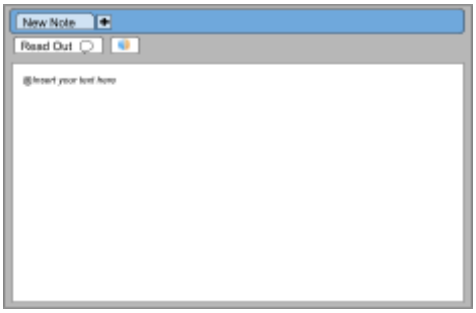
- **Emotion Recognition Metrics:** This table will provide details of the emotion recognition accuracy, precision, recall, and F1 score.

- **Speech Emotion Conversion Output:** This table will provide details of the final speech output after adjusting it to reflect the appropriate emotion.

Report Layout: The actual report layout will be put into an appendix and will include the above tables in a tabular format. The report will also include graphs and charts to visualize the data, such as a bar chart to show the emotion recognition accuracy, precision, recall, and F1 score. The report will be structured in a logical flow, starting

from the input data, proceeding through the emotion detection and speech emotion detection models, and ending with the speech emotion model's output.

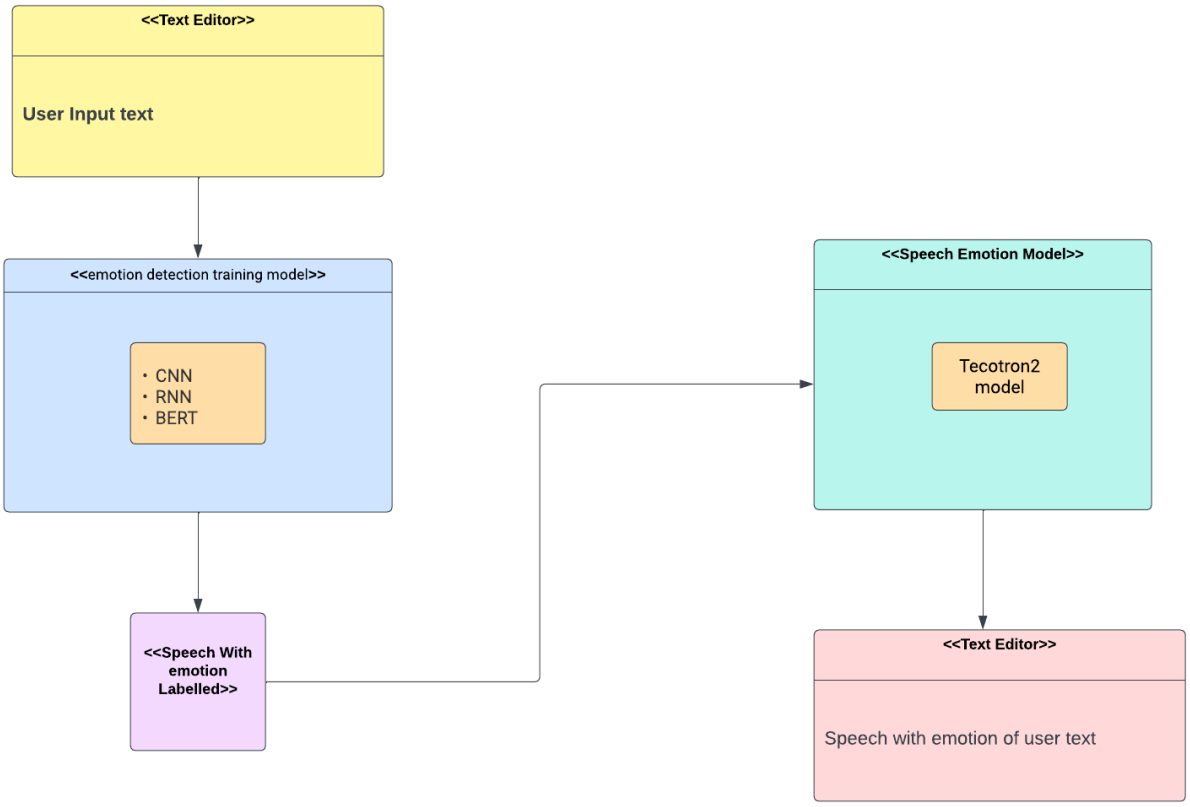
9. External Interfaces



Text Input From the user

Speech is generated with emotion

10. Packaging and Deployment Diagram



11. Help

User Manual: A user manual will be provided as a PDF document. This document will provide detailed instructions on how to use the system, including step-by-step guides and screenshots.

Technical Manual: A technical manual will be provided as a PDF document. This document will provide detailed information about the system architecture, software and hardware requirements, installation and configuration instructions, and troubleshooting guides.

Tutorials: The system will include tutorials to help users learn how to use the system. These tutorials will be available in the form of videos and step-by-step guides.

12. Design Details

a. **Platforms:** The project will be developed using Python programming language and will primarily run on desktop or server environments with appropriate hardware and software requirements. The speech emotion recognition and conversion models will utilize libraries such as Tensorflow, Keras, PyTorch, and Librosa for processing and analysis.

b. **Systems:** The project will require access to labeled speech and text data, which can be sourced from various datasets or collected through crowd-sourcing platforms. Additionally, the project may require access to hardware resources such as GPUs for training deep learning models.

c. **Processes:** The project will involve several processes such as data collection and preprocessing, training and tuning of the emotion recognition and speech conversion models, and deployment of the final model for production use. The project may also require ongoing maintenance and updates to ensure optimal performance and reliability.

d. **Novelty:** The project aims to address the challenge of converting neutral speech to emotional speech, which is a relatively new and evolving field in natural language processing and speech technology. The use of deep learning models and innovative approaches such as DSCNN for speech emotion recognition and conversion can provide significant advancements in this area.

e. **Interoperability:** The project will be designed to be interoperable with various platforms and systems that may require the use of emotional speech, such as virtual assistants, chatbots, and voice-enabled applications.

f. **Reliability:** The project will aim to achieve high reliability by testing and validating the system through rigorous quality assurance processes. The system will be designed to handle errors and exceptions gracefully and provide robust and reliable output.

g. **Maintainability:** The project will be designed for easy maintenance and updates, with clear and modular code architecture, appropriate documentation, and version control systems. This will enable easy troubleshooting and debugging of issues and ensure long-term sustainability of the system.

h. **Portability:** The project will aim to be portable and platform-independent, with the ability to run on various operating systems and hardware configurations. This will enable easy deployment and use of the system across different environments.

i. **Legacy to Modernization:** The project will aim to modernize speech technology by incorporating cutting-edge deep learning techniques and innovative approaches for speech emotion recognition and conversion. This can provide significant advancements in the field and contribute to the development of more advanced and intelligent speech systems.

j. **Application Compatibility:** The project will aim to be compatible with various applications and platforms that may require the use of emotional speech. This can include text editors, virtual assistants, chatbots, and other voice-enabled applications.

k. **Resource Utilization:** The project will aim to maximize resource utilization by optimizing hardware and software resources for efficient processing and analysis of speech and text data. This will enable high-performance and cost-effective implementation of the system.

l. **Security:** The project will implement appropriate security measures to protect sensitive data and prevent unauthorized access or misuse of the system. This may include encryption of data during transmission and storage, user authentication and access control, and secure deployment of the system.

Appendix A: Definitions, Acronyms and Abbreviations

Emotion detection model: A model that analyzes text or speech and identifies the corresponding emotion, such as anger, sadness, happiness, or sarcasm.

Preprocessing: The process of cleaning and preparing input data for further processing.

Stop words: Commonly used words that are often removed from text data during preprocessing.

Explicit words: Words that may be inappropriate or offensive in certain contexts.

Punctuation: Marks such as periods, commas, and quotation marks that are used to separate and structure text.

Deep learning: A subset of machine learning that involves training neural networks to perform complex tasks.

Sarcasm: A form of verbal irony in which the intended meaning of a word or phrase is the opposite of its literal or expected meaning.

UK/US accent: A way of pronouncing English that is specific to either the United Kingdom or the United States.

Annotated text: Text that includes additional information, such as labels or tags, to provide context and meaning.

Role-based dialogue format: A format of text that is used in a dialogue between two or more people, with each person assigned a specific role.

GPU: Graphics Processing Unit, a type of computer processor that is optimized for parallel processing and commonly used for machine learning applications.

Prosody: The patterns of stress and intonation in speech that convey emotional content.

NLP - Natural Language Processing
GPT - Generative Pre-trained Transformer
LSTM - Long Short-Term Memory
CNN - Convolutional Neural Network
RNN - Recurrent Neural Network
DSCNN - Deep Stridal CNN
POS - Part-of-Speech
SVD - Singular Value Decomposition
PCA - Principal Component Analysis
GUI - Graphical User Interface
UI - User Interface
MVP - Minimum Viable Product

Appendix B: References

[1] X. Cai, D. Dai, Z. Wu, X. Li, J. Li and H. Meng, "Emotion Controllable Speech Synthesis Using Emotion-Unlabeled Dataset with the Assistance of Cross-Domain Speech Emotion Recognition," ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toronto, ON, Canada, 2021, pp. 5734-5738, doi: 10.1109/ICASSP39728.2021.9413907

[2] Effective Text Data Preprocessing Technique for Sentiment Analysis in Social Media Data Saurav Pradha School of Computing and Mathematics Charles Sturt University Melbourne, Victoria, Senior Australia saurav.pradha54@gmail.com Malka N. Halgamuge Member, IEEE Dep. of Electrical and Electronic Engineering The University of Melbourne Victoria 3010, Australia malka.nisha@unimelb.edu.au Nguyen Tran Quoc Vinh Faculty of Information Technology The University of Da Nang - University of Science and Education, Vietnam ntquocvinh@ued.udn.vn.

[3] Wang, Yuxuan, R. J. Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J. Weiss, Navdeep Jaitly, Zongheng Yang, Ying Xiao, Z. Chen, Samy Bengio, Quoc V. Le, Yannis Agiomyrgiannakis, Robert A. J. Clark and Rif A. Saurous. “Tacotron: Towards End-to-End Speech Synthesis.” Interspeech (2017).

[4] P. Chandra et al., "Contextual Emotion Detection in Text using Deep Learning and Big Data," 2022 Second International Conference on Computer Science, Engineering

and Applications (ICCSEA), Gunupur, India, 2022, pp. 1- 5, doi: 10.1109/ICCSEA54677.2022.9936154.

[5] Ren, Yi, Ruan, Yangjun, Tan, Xu, Qin, Tao, Zhao, Sheng, Zhao, Zhou, and Tie Liu. "FastSpeech: Fast, Robust and Controllable Text to Speech." ArXiv, (2019). Accessed February 10, 2023. https://doi.org/10.48550/arXiv.1905.09263.

[6] Valentini-Botinhao, C., Wang, X., Takaki, S., Yamagishi, J. (2016) Investigating RNN-based speech enhancement methods for noise-robust Text-to-Speech. Proc. 9th ISCA Speech Synthesis Workshop, 146-152.

[7] S. Furui, T. Kikuchi, Y. Shinnaka and C. Hori, "Speech-to-text and speech-to-speech summarization of spontaneous speech," in IEEE Transactions on Speech and Audio Processing, vol. 12, no. 4, pp. 401-408, July 2004,

[8] Text to Speech Conversion with Emotion Detection Article in International Journal of Applied Engineering Research · January 2018 Srinivasan Rajendran SRM Institute of Science and Technology 19 PUBLICATIONS 66 CITATIONS

[9] A Comprehensive Review of Speech Emotion
Recognition SystemsTAIBA MAJID WANI 1, TEDDY SURYA GUNAWAN 1,3, (Senior Member, IEEE), SYED ASIF AHMAD QADRI 1, MIRA KARTIWI 2, (Member, IEEE), AND ELIATHAMBY AMBIKAIRAJAH 3, (Senior Member, IEEE)

[10] SPEECH EMOTION RECOGNITION USING DEEP NEURAL NETWORK CONSIDERING VERBAL AND NONVERBAL
SPEECH SOUNDS Kun-Yi Huang, Chung-Hsien Wu, Qian-Bei Hong, Ming-Hsiang Su and Yi-Hsuan Chen, Department of Computer Science and Information Engineering, National Cheng Kung University, Tainan, Taiwan

[11] Carrillo-de-Albornoz, Jorge & Plaza, Laura & Gervás, Pablo. (2012). SentiSense: An easily scalable concept-based affective lexicon for sentiment analysis.

[12] Huan, Jeow & Sekh, Arif Ahmed & Quek, Chai & Prasad, Dilip. (2022). Emotionally charged text classification with deep learning and sentiment semantic. Neural Computing and Applications. 34. 10.1007/s00521-021-06542-1.

[13] U. Rashid, M. W. Iqbal, M. A. Skiandar, M. Q. Raiz, M. R. Naqvi and S. K. Shahzad, "Emotion Detection of Contextual Text using Deep learning," 2020 4th International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT), Istanbul, Turkey, 2020, pp. 1-5, doi: 10.1109/ISMSIT50672.2020.9255279.

[14] U. Rashid, M. W. Iqbal, M. A. Skiandar, M. Q. Raiz, M. R. Naqvi and S. K. Shahzad, "Emotion Detection of Contextual Text using Deep learning, " 2020 4th International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT), Istanbul, Turkey, 2020, pp. 1-5, doi: 10.1109/ISMSIT50672.2020.9255279.

[15] A Comprehensive Review of Speech Emotion
Published in: IEEE Access (Volume: 9)
Page(s): 47795 - 47814
Date of Publication: 22 March 2021
Electronic ISSN: 2169-3536
INSPEC Accession Number: 20965838
DOI: 10.1109/ACCESS.2021.3068045
Publisher: IEEE

[16] Speech Emotion Recognition using Convolution Neural Networks and Deep Stride Convolutional
Neural Networks - 2020
Published in: 2020 6th International Conference on Wireless and Telematics (ICWT)
Date of Conference: 03-04 September 2020
Date Added to IEEE Xplore: 03 November 2020
ISBN Information:
INSPEC Accession Number: 20133021
DOI: 10.1109/ICWT50448.2020.9243622

[17] Demszky, Dorottya, et al. "GoEmotions: A dataset of fine-grained emotions." arXiv preprint
arXiv:2005.00547 (2020).

[18] Text to Speech Conversion with Emotion Detection Article in International Journal of Applied Engineering

Research · January 2018

Appendix C: Record of Change History

[This section describes the details of changes that have resulted in the current High-Level Design document.]

#	Date	Document Version No.	Change Description	Reason for Change
1.				
2.				
3.				