RSET
RAJAGIRI SCHOOL OF
ENGINEERING & TECHNOLOGY
(AUTONOMOUS)

Project Phase 2 Report On

Emotive Malayalam Text-to-Speech

Submitted in partial fulfillment of the requirements for the award of the degree of

Bachelor of Technology

in

Computer Science and Engineering

By

Akash Vijay (U2003016)
Aleena Mary Karatra (U2003023)
Alvin George Viji (U2003029)
Ashly Sabu (U2003044)

Under the guidance of

Ms. Sherine Sebastian

Department of Computer Science and Engineering
Rajagiri School of Engineering & Technology (Autonomous)
(Parent University: APJ Abdul Kalam Technological University)
Rajagiri Valley, Kakkanad, Kochi, 682039
April 2024

CERTIFICATE

This is to certify that the project report entitled ***"Vikara: An Emotional Malayalam Text-to-Speech Synthesizer"*** *is a bonafide record of the work done by* ***Akash Vijay (U2003016)***, ***Aleena Mary Karatra (U2003023)***, ***Alvin George Viji (U2003029)***, ***Ashly Sabu (U2003044)*** *submitted to the Rajagiri School of Engineering & Technology (RSET) (Autonomous) in partial fulfillment of the requirements for the award of the degree of Bachelor of Technology (B. Tech.) in Computer Science and Engineering during the academic year 2023-2024.*

Ms. Sherine Sebastian
Asst. Professor
Dept. of CSE
RSET

Dr. Sminu Izudheen
Professor
Dept. of CSE
RSET

Dr. Preetha K G
Head of the Department
Professor
Dept. of CSE
RSET

ACKNOWLEDGMENT

We wish to express our sincere gratitude towards **Dr. P S Sreejith**, Principal of RSET, and **Dr. Preetha K G**, Head of the Department of Computer Science for providing us with the opportunity to undertake the project, "Emotive Malayalam Text-to-Speech".

We are highly indebted to our project coordinator, **Dr. Sminu Izudheen**, Professor, for her valuable support.

It is indeed our pleasure and a moment of satisfaction to express our sincere gratitude to our project guide **Ms. Sherine Sebastian** for her patience and all the priceless advice and wisdom she has shared with us.

Last but not the least, we would like to express our sincere gratitude towards all other teachers and friends for their continuous support and constructive ideas.

Akash Vijay
Aleena Mary Karatra
Alvin George Viji
Ashly Sabu

Abstract

The synthesis of emotionally expressive speech has gained significant attention in recent years, as it plays a crucial role in improving the naturalness and effectiveness of human-computer interaction. The project focuses on developing an Emotive Malayalam Text-to-Speech (TTS) system, aiming to infuse emotive qualities into synthesized speech in Malayalam language. The proposed system uses natural language processing and deep learning techniques to capture and reproduce the exact emotions present in written text.

The project involves the creation of a large and diverse emotional speech corpus in Malayalam, incorporating a spectrum of emotions such as happiness, sadness, anger, and surprise. Deep neural networks, including recurrent and attention-based models, are employed to train the TTS system, allowing it to learn the intricate patterns and variations associated with different emotional states.

The evaluation of the Emotive Malayalam TTS system involves subjective and objective measures, including perceptual listening tests and emotion classification accuracy. The outcomes of the project not only contribute to the advancement of emotional speech synthesis in Malayalam languages but also have practical applications in various domains such as assistive technology, virtual assistants, and interactive entertainment.

Contents

List of Abbreviations

TTS - Text to Speech

MFCC - Mel-frequency cepstral coefficients

ANN - Artificial Neural Network

ALO - Ant Lion Optimizer

DNN - Deep Neural Network

MLP - Multilayer Perceptron

RAVDESS - Ryerson Visual Audio Data Database

FFT - Fast Fourier Transform

DFT - Discrete Fourier Transform

ResNet - Residual Networks

HNR - Harmonic to Noise Rate

ZCR - Zero Crossing Rate

TEO - Teager Energy Operator

SVM - Support Vector Machine

CNN - Convolutional Neural Network

GRU - Gated Recurrent Unit

SMO - Sequential Minimal Optimization

LSTM - Long Short-Term Memory

RNN - Recurrent Neural Network

NLP - Natural Language Processing

SNA - Social Network Analysis

LDA - Latent Dirichlet Allocation

POS - Parts of Speech

API - Application Programming Interface

MSSA - Modified Salp Swarm Algorithm

SSA - Salp Swarm Algorithm

HDL-SA - Hybrid Deep Learning-Sentimental Analysis

ABSA - Aspect-Based Sentiment Analysis

RT - Retweet Removal

TAN - Transformer Attention Network

ROC - Receiver Operating Characteristic

List of Figures

List of Tables

Chapter 1

Introduction

The project addresses speech synthesis technology by concentrating on the development of an Emotive Malayalam Text-to-Speech (TTS) system. While current TTS systems excel in delivering intelligible speech, the incorporation of emotional expressiveness remains a challenging frontier crucial for enhancing user engagement. Focused on the Malayalam language, spoken by millions in Kerala, India, in which the project employs advanced natural language processing and deep learning techniques to create a TTS system that not only accurately reproduces Malayalam's linguistic features but also captures the subtle emotional nuances present in written text. By utilizing translation models and deep neural networks, including recurrent and attention-based models, the research aims to enable the TTS system to reproduce a spectrum of emotions. Integrated sentiment analysis tools dynamically adjust prosody, intonation, and pacing based on the emotional content of the input text. The significance of the project extends beyond technical advancements, contributing to the broader goal of making human-computer interaction more natural, intuitive, and emotionally resonant.

1.1 Background

The history of emotive text-to-speech (TTS) technology reflects the ongoing evolution of artificial intelligence, natural language processing, and the quest to create more human-like and emotionally expressive synthetic speech. While the earliest TTS systems primarily focused on delivering intelligible and coherent speech, the integration of emotional cues became a significant research pursuit in the late 20th century and gained more prominence in the 21st century.

- Early Development (1950s-1980s): The earliest TTS systems emerged in the 1950s and were primarily designed to convert text into speech for basic information dis-

semination. These systems lacked emotional expressiveness and focused on clarity and comprehensibility. Through the 1980s, TTS research concentrated on improving the overall quality and naturalness of synthesized speech.

- Prosodic and Emotional Cue Research (1990s): In the 1990s, researchers began to explore the role of prosody—intonation, rhythm, and stress patterns—in conveying emotion in speech. Efforts were made to incorporate these prosodic features into TTS systems to make synthesized speech sound more natural. However, the incorporation of explicit emotional cues was still limited during this period.

- Emotional Speech Synthesis (2000s): The 21st century witnessed a more dedicated focus on emotional speech synthesis. Researchers started investigating the integration of emotion-specific models and databases to infuse synthesized speech with a range of emotions, such as happiness, sadness, anger, and surprise. This era saw advancements in both the modeling of emotional features and the availability of emotional speech corpora for training TTS systems.

- Deep Learning and Neural Networks (2010s-present): The advent of deep learning and neural network technologies revolutionized the field of TTS. Deep neural networks, including recurrent neural networks (RNNs) and attention mechanisms, enabled more sophisticated modeling of linguistic and emotional features. These technologies allowed TTS systems to capture subtle nuances in emotional expression and deliver more natural-sounding speech.

1.2 Problem Definition

There is a distinct lack of exploration in the field of Text-to-Speech Synthesis for the Malayalam language. Existing TTS systems designed for Malayalam lack the capability to convey emotions effectively in the synthesized speech output. The project aims to develop an emotive Malayalam text-to-speech synthesizer that helps to produce an emotive audio output for the Malayalam input text given.

1.3 Scope and Motivation

The primary scope involves developing an Emotive Malayalam Text-to-Speech (TTS) Synthesizer to significantly enhance emotional expressiveness in synthesized speech. Naturalness in speech infusing emotional tones into speech synthesis, creating a more natural and engaging auditory experience. It can improve accessibility for individuals with visual impairments, making it easier for them to consume content online or interact with digital devices in a more emotionally resonant manner. Helps in promoting the Malayalam language and culture by providing a tool for generating expressive and emotionally rich content in the native language.

The project is motivated by the recognition of the current limitations in existing Malayalam TTS systems, particularly the lack of emotion in synthesized speech. Humanizing human-computer interaction by making synthetic speech more emotionally resonant, improving the overall user experience. The goal is to make synthetic speech sound more human-like, allowing users to interact with devices, applications, or virtual assistants in a way that feels natural and relatable. By focusing on the Malayalam language, the project aims to cater to the cultural and linguistic needs of the Malayalam-speaking population.

1.4 Objectives

1. To develop a Text-to-Speech Synthesizer for the Malayalam Language.
2. To analyze the emotion of the input text using sentiment/emotion analysis
3. To generate emotion in the speech using emotion voice conversion after identifying the emotion in the text.
4. To better the lives of the visually impaired, by providing expressive and emotive synthesized speech for Malayalam audiobooks, essays, and so on.
5. To develop the ability to adapt the emotional tone based on the context of the text. Understand the content and adjust the emotional expression accordingly for a more coherent and contextually appropriate output.

1.5 Challenges

Malayalam is a language rich in emotions and expressions. Capturing and representing these nuances in TTS can be complex. Achieving natural and emotive voice quality in Malayalam can be challenging due to the unique phonetic characteristics of the language. Poor voice quality can result in an unnatural or robotic-sounding output, diminishing the effectiveness of the TTS system. Building emotional corpora (datasets) for Malayalam may be limited compared to languages with extensive resources so the lack of a diverse emotional dataset may hinder the training process and result in a less accurate emotional representation. Hence, utilizing translation software, or training the models using a high resource language like English, would address this challenge, while bringing concerns of its own. Translation and training the model using a high resource language may affect the accuracy of the model when it comes to identifying the emotion present in the language, as well as converting the emotion present in the neutral speech. A significant challenge is of assuring the accuracy of the translation model, or the integrity of the text may be lost causing the emotion predicted to be inaccurate.

1.6 Assumptions

1. The TTS synthesizer is assumed to accurately convert Malayalam text into speech, considering correct pronunciation and linguistic nuances. The project's effectiveness relies on the assumption that the TTS system can faithfully reproduce Malayalam speech from diverse textual inputs.

2. The sentiment/emotion analysis module is assumed to reliably recognize and categorize emotions within the input text. The input text is initially translated to English which is followed by the identification of the emotion. The accuracy of emotion voice conversion depends on the correct identification of emotions as well as the accuracy of the translation, and any inaccuracies in the analysis may impact the emotive quality of the synthesized speech.

3. The project assumes a universal emotional expression model that can be applied effectively to Malayalam language-specific emotional cues. The emotional voice conversion relies on the assumption that emotions are expressed in a generally con-

sistent manner across different languages, allowing for the application of a model developed for Malayalam, but trained on a high resource language like English.

4. The project assumes that the emotion synthesis process is culturally sensitive and aligns with the emotional expressions common in Malayalam culture. Cultural nuances play a vital role in emotional expression, and the assumption is that the project considers these nuances for an authentic emotional representation.

1.7 Social/Industrial Relevance

1. **Accessibility for Visually Impaired:** The project improves accessibility for the visually impaired by using expressive synthesized speech, enhancing engagement with written content.

2. **Inclusive Educational Resources:** Tailored synthesized speech for audiobooks and essays promotes inclusive education, ensuring equitable access for visually impaired individuals.

3. **Emotional Connection in Communication:** The project seeks to enhance synthesized speech, creating a more emotionally connected communication experience for visually impaired users.

4. **Empowerment Through Technology:** The project uses technology to empower visually impaired individuals in navigating a text-dependent world.

1.8 Organization of the Report

The organization of the paper is as follows:

• Introduction- Gives a brief overview of the reason why the project is being done, scope, motivation, and objectives.

• Literature Survey- Explains about different research papers used to achieve goals for the project. And gives a detailed summary of the different methodologies used in the research paper.

• Requirements- Information about the hardware and software that will be required to successfully complete the project.

• System Architecture- Gives a brief system overview and the proposed system architecture including models.

• System Implementation- Describes the datasets used, proposed methodologies, algorithms, user interface design, and other such information to implement project.

• Results and Discussions- Gives a brief overview of the entire project and how successful the project was along with analyzing the results obtained.

• Conclusions and Future Scope- This section concludes the overview of the work and discussion on the future applications.

• References

• Appendix

1.9 Conclusion

In conclusion, the chapter provides a succinct overview of the Emotive Text to Speech project, elucidating its background, personal motivations, objectives for success, and common challenges encountered. It underscores the societal and industrial relevance of the endeavor, emphasizing its potential to enhance human-computer interaction and foster more empathetic digital experiences across various domains. By addressing these fundamental aspects, the project endeavors to redefine communication paradigms, offering a glimpse into a future where artificial systems seamlessly integrate emotional intelligence into synthesized speech, enriching interactions and bridging the gap between technology and human emotion.

Chapter 2

Literature Survey

2.1 An efficient adaptive artificial neural network based text-to-speech synthesizer for the Hindi language

The paper, An efficient adaptive artificial neural network based text to speech synthesizer for the Hindi language suggests a new methodology for designing an effective context-free high-quality speech artificially generated from written script. using MFCC features, linguistic constraints, and an ANN model based on the ALO algorithm. The paper gives an overview of different approaches taken to develop speech synthesizers. Existing frameworks and models for speech synthesis in the case of Hindi are also considered herein.

Objective measures such as prediction error, standard deviation, and linear correlation coefficient are used to evaluate the proposed approach. Experimental results suggest that the proposed ALO-ANN approach is more accurate in prediction than other models such as deep neural networks (DNN) and traditional ANNs. With applications in areas such as accessibility, education, communications, and entertainment the new text-to-speech synthesizer can help advance Hindi-speaking society's technology.

The only problem between this paper and sisters is that they all talk about validation, generalization, ethical considerations, benchmarking, and user-oriented evaluation. Addressing these issues would strengthen the paper's impact on speech processing and related areas. All in all, the paper outlines a new way of building such a text-to-speech synthesizer for Hindi. Its fields of application are many and include improving accessibility or usability as well as learning foreign languages; it also has applications in assistive communication systems that help people with impaired hearing (either temporarily through injury or permanently due to disease), and multimedia applications like online education.

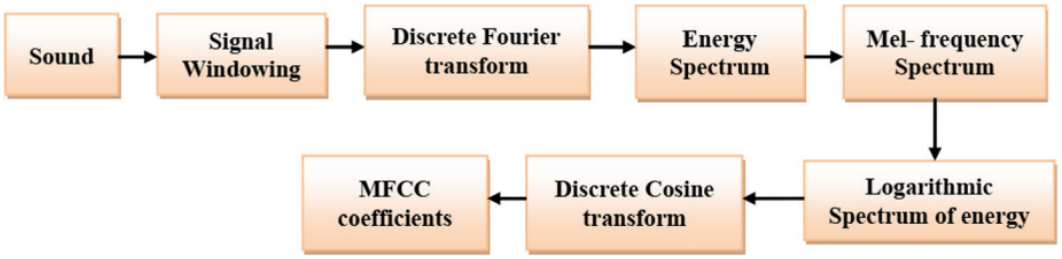


Figure 2.1: Block diagram of the MFCC feature extraction process

2.2 Classification of emotions from speech signals using machine learning

The difficult task of recognizing emotions from speech signals is thoroughly explored in the study "Classification of Emotions from Speech Signal Using Machine Learning". The authors discuss the challenge of precisely obtaining emotion-oriented speech features and suggest that Mel Frequency Cepstral Coefficients (MFCC) be used in feature extraction. They use two separate datasets, RAVDESS and a Malayalam database made from movie clips, to show how well a multilayer perceptron (MLP) performs in classifying speech signals based on the extracted features. They achieve recognition accuracy of 78.56 percent and 84.3 percent, respectively.

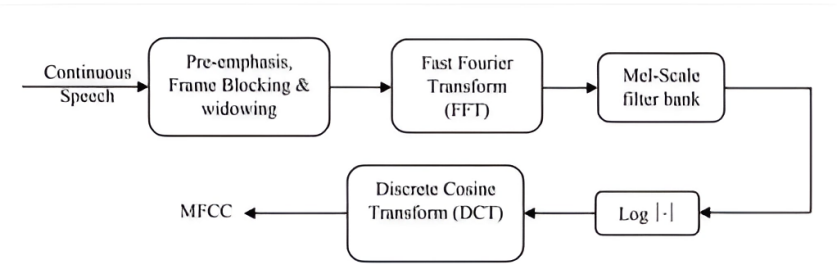


Figure 2.2: Block diagram of MFCC extraction

The study highlights the significance of efficient data structures, algorithms, and software engineering tools in the implementation of a real-time system for emotion classification. The potential applications of this research include human-computer interaction, medical management, tutoring, defense, and safety. The findings emphasize the importance of accurate feature extraction for successful emotion classification from speech sig-

nals using machine learning techniques.

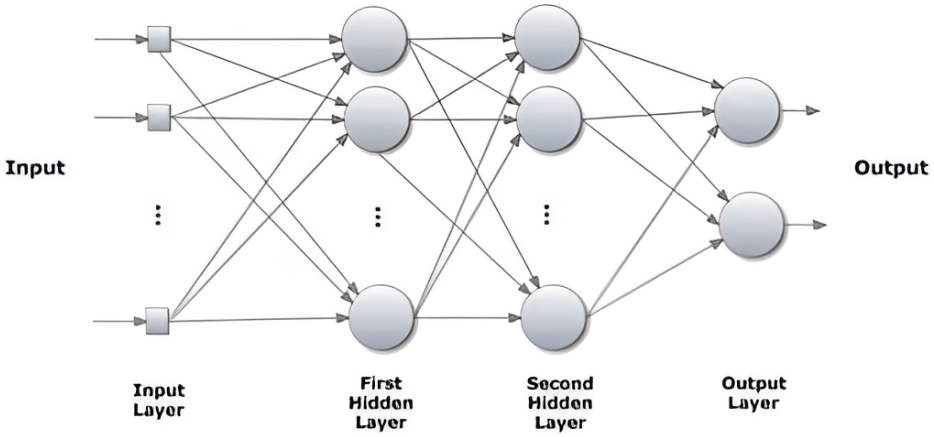


Figure 2.3: Multilayer perceptron

2.3 Sentimental analysis using hybrid deep learning and optimization technique

2.3.1 Overview

The study discusses an efficient sentiment analysis technique for Twitter using a hybrid deep learning and optimization approach. Sentiment analysis is the process of identifying and classifying opinions expressed on social networks such as Twitter. The proposed system uses Parts of Speech (POS) tagging and hybrid deep learning techniques to analyze tweets and classify them into five sentiment categories. The results show that the proposed approach achieves better accuracy and efficiency compared to other commonly used models such as random forest, naive base and decision tree classifier.

2.3.2 System architecture

The model uses a neural network for analyzing sentiment in the Twitter data. The neural network uses tensor flow and bag words model. The proposed solution uses Twitter API for fetching the data and then preprocessing the data before it passes to the ANN for classification. The system overview is shown in the figure 2.4.

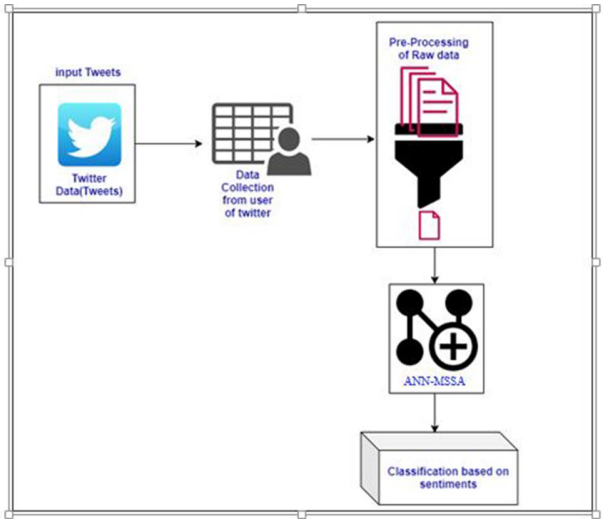


Figure 2.4: HDL-SA technique.

2.3.3 Hybrid deep learning technique

This section contains the detailed description of the Hybrid deep learning technique from the data preprocessing to the training of the model.

1. Dataset description

   This study utilizes the updated version of the openly available SemEVAL-2017 dataset with 5 5-point ordinal scale. The dataset contains more than 20 thousand tweets and the tweet sentiment score is identified for each tweet. The semantic orientation and its equivalent sentiment score orientation ranging from highly positive to highly negative on a five point scale based on the sentiment score is mapped as in table 2.1.

Table 2.1: Semantic orientation and its equivalent sentiment score

S. no	Semantic Orientation	Sentiment Score
1	Highly Positive 1	2
2	Positive	1
3	Neutral	0
4	Negative	-1
5	Highly Negative	-2

2. Data pre-processing

Data pre-processing involves several steps to extract important information from raw tweets. This includes converting tweets to non-case sensitive, tokenization, excluding usernames and hashtags, eliminating stop words, removing non-alphabetic content, lemmatization, and merging tokens into sentences. This process helps to clean the data and extract the information required for sentiment analysis. The step-by-step detailed diagram is shown in Figure 2.5.

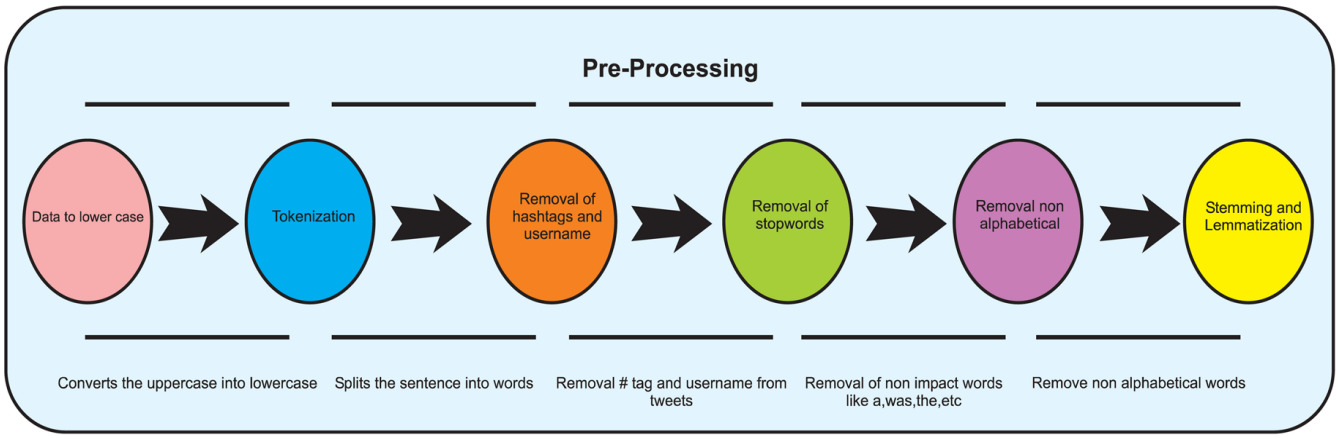


Figure 2.5: Pre-processing of data.

3. Sentimental analysis using artificial neural network (ANN)

The step-by-step process for forwarding to the neural network is given below. The ANN structure can be pictured in beneath Figure 3.3.

- Step 1: Initialization of Input Parameters - Initialize the input parameters as a set: $G_i = \{G_1, G_2, G_3, \ldots, G_n\}$.
- Step 2: Hidden Layer Configuration - Define a hidden layer containing neurons labeled $HD1, HD2, \ldots, HDn$. - These neurons play a crucial role in connecting the hidden layers to the output layers.
- Step 3: Start of Weight - Introduce the weights of the input layer (α_m) and the hidden layer (β_{mn}), specifying that they are initialized in this layer.
- Step 4: Activation Task Evaluation - Present equation (1), which evaluates the activation function of artificial neural network (ANN). - In this equation,

HD_{basis} is denoted as the basis function, and HD_{acti} is denoted as the activation function. - In particular, $HD_{basis} = \sum_{j=1}^{N} G_i \times \beta_{ij}$. - The activation function is calculated as $HD_{acti} = \sum_{x=1}^{l} \frac{1}{1+\exp(-H_b)}$.

- Step 5: Fitness Function Development - Describe the derivation of a fitness function designed to predict the performance of the proposed model. - Emphasizing the objective of achieving maximum accuracy in forecasts.
- Step 6: Fitness Value Determination - Expand the process of determining the fitness value for each solution. - Specify that the optimal fitness is selected based on the maximum accuracy criterion. Formally, $F_{func} = \max |accuracy|$.

4. Modified Salp swarm algorithm (MSSA)

The study presents a Modified Salp Swarm Algorithm (MSSA) for weight optimization in sentiment analysis. The Salp Swarm Algorithm (SSA) is a metaheuristic algorithm inspired by the swarming behavior of marine organisms called salps. The purpose of SSA is to find the optimal location in the search space, similar to how salps search for food in the ocean. The proposed MSSA is used to update the weights of an Artificial Neural Network (ANN) for sentiment analysis. It simulates the behavior of a salp chain in finding the optimal position in the search space. This article discusses the process of updating the positions of leaders and followers in a swarm, as well as the application of Newton's law of motion to update the positions of followers. The results indicate that MSSA effectively optimizes the weights of ANN, leading to improved accuracy and reduced time consumption in sentiment analysis. The steps are described below.

- Swarm Structure and Objective: - The position of the salps is stored in the matrix S, and the food source f_s is established as the goal of the swarm. - The weights of the Neural Network (NN) are embedded in the Salp Swarm Algorithm (SSA) population.
- Leader Status Update: - The leader position (S_j^1) is updated based on the food source using the conditions:

- Parameter Definitions: $S1_j$ indicates the position of the first salp (leader) in the jth dimension. f_j represents the position of the food source in the jth dimension. U_j and L_j are upper and lower bounds; R_1, R_2, and R_3 are random numbers.
- Significance of R_1 in SSA: R_1 is a critical parameter in SSA, adjusting exploration and exploitation, defined as $R_1 = \frac{2e(4l)}{L}$, where l is the current iteration and L is the maximum number of iterations.
- Follower Status Update: Newton's law of motion is applied to update the position of the follower salps (S_j^i), simulating salps chains:

$$S_j^i = \frac{S_j^i + S_j^{(i-1)}}{2}$$

- Navigation and Maintenance: Updating food sources during navigation is important for optimal positioning in the ever-changing salt chain.
- Optimal Weight Determination: When the optimal salt level increases, the optimal weights are determined, which separate the ideal weights for the output and hidden layers of the NN.
- Neural Network Structure: A hidden neuron layer connects neurons to the output layer, and cognitive test modeling is applied to increase the robustness of the simulation.

5. Result and discussion

The Results and Discussion section presents a comparison of the proposed hybrid deep learning technique with existing classifiers, such as decision trees, random forests, and naive bases. Performance measures evaluated include accuracy, precision, recall, and F-measure. The proposed approach using Artificial Neural Network (ANN) shows significantly higher accuracy, precision, recall, and F-measure compared to existing classifiers. Additionally, the section includes a detailed analysis of the time taken in terms of the number of tweets for each classifier. The results show that the proposed ANN classifier outperforms the existing classifiers in terms of time efficiency. These findings suggest that the hybrid deep learning approach is more

effective in sentiment analysis for Twitter data. In Table 2.2, the proposed HDL-SA technique with ANN classifier is compared with existing classifiers are decisions tree, random forest, and Naive Bayes classifiers.

Classifier	Accuracy (%)	Precision (%)	Recall (%)	F-measure
Decision Tree	73	73.2	71.2	0.7012
Random Forest	78	88.8	85.5	0.854
Naive Bayes	70	54.4	70.4	0.741
ANN	92	91.3	93.1	0.9312

Table 2.2: Classifier Performance Metrics

2.4 MASS: Multi-Task Anthropomorphic Speech Synthesis Framework

As discussed earlier, this model is constituted of three different modules connected in series: TTS module, DEVC, and DSVC. As each module is separately trained, the model uses separate single-task datasets for training. This interconnected structure allows for separate training of modules, preventing interference and maintaining the flexibility to replace modules for enhanced synthesis effects.

The TTS takes the input speech and generates a neutral speech as a normal TTS model does. The output of the TTS, which is the generated speech, is fed into the DEVC module, which has a structure similar to that of CycleGAN, that incorporates the required emotion into the generated speech. The speech which now has emotional features, is fed to the DSVC module, which now performs the speaker identity conversion, to produce the final output, a speech with the required identity and emotional characteristics.

It should be noted that the DEVC is trained using a dataset with speech uttered by a single speaker, hence ensuring that identity characteristics are preserved during the emotion voice conversion process. Hence, the positions of the DEVC and DSVC modules cannot be interchanged in the serial connection. As discussed above, the TTS module generates the speech from the input text as shown in the diagram. This neutral speech, is fed into the DEVC module that extracts its audio features, including the Spectrogram

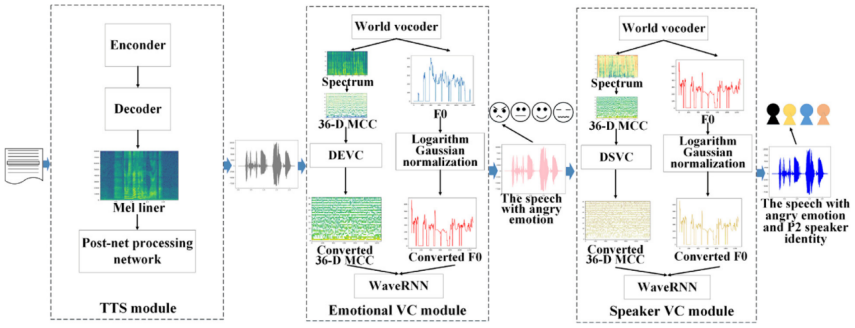


Figure 2.6: Architecture Diagram of MASS Model

and the fundamental frequency F0, to manipulate this and add the emotion into the audio. The 4 expressions denote the 4 emotions incorporated- angry, neutral, amused, and disgusted. This speech is given as input to the DSVC module that converts the speaker identity from neutral to a specific identity, similarly, as shown with the 4 different colors denoting the 4 different identities.

The model uses WaveRNN as a vocoder to reconstruct the speech from the spectral features after modification from the DSVC module. To ensure that the emotional and identity characteristics are not lost during the reconstruction, the F0 and Mel-Cepstral Coefficient(MCC) features are concatenated to the data.

2.4.1 Text-to-Speech Module

The MASS method uses a TTS transformer model to synthesize the neutral speech for conversion from the input text. This model is known for its efficiency and performance standards and is an improved version of the Tacotron2 model discussed earlier. This TTS model uses a multi-head attention mechanism to replace the RNN and attention-mechanism model in the Tacotron2 model for better performance. However, there is a trade off of advantages between choosing a multi-head attention mechanism versus a self-attention mechanism.

2.4.2 Emotion Voice Conversion Module

The DEVC is and integral constituent of the MASS model that performs the task of incorporating emotion into neutral sounding text. The architecture of the same is as

illustrated below:

The Spectral information of the generated neutral speech is extracted using the WORLD

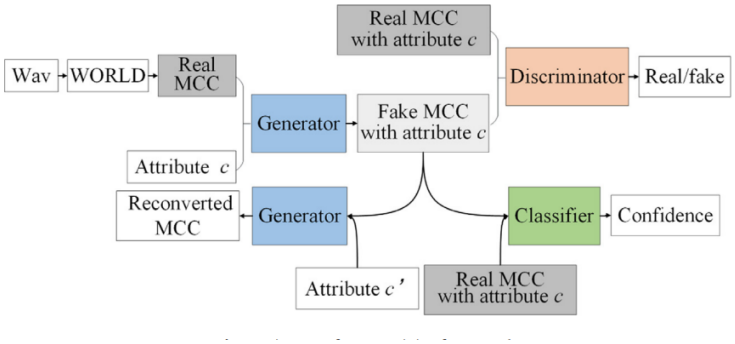


Figure 2.7: Architecture Diagram of MASS Model

vocoder, thus obtaining the real MCC, or the MCC of the generated speech. This, along with the attribute c, or the required emotion to be incorporated, is fed as input to the generator. The generator generates a new fake MCC that has the emotion, or attribute c. Both these fake and real audio samples with the attribute c are fed into the discriminator, whose purpose is to correctly identify the real and fake audio samples.

This generated emotional audio data is also sent to the classifier that aims to classify it into the correct emotional category. Both the discriminator and classifier act as measures of the generator's performance. The second generator, re-generates the audio sample using the previous neutral attribute as a reference, to ensure that there is not too much deviation from the original audio sample as to make the generated speech sound unnatural. Hence, there are three adversarial networks in the DEVC module:

1. The Generator-Discriminator Network

   The loss function of the discriminator is its training objective that enables it to learn and discriminate between the audio samples generated, and real audio samples. The goal of the generator is to maximize that loss so that it generates audio that is indistinguishable from real audio.

2. The Generator-Classifier Network

   The generator works similarly to before, except that here the goal of the generator is to generate the emotion that is discernible and which is correctly classified by the classifier. The classifier's equation defines the optimal quantitatives of the classifier

so that it correctly identifies and classifies each emotion.

3. The Generator-Generator Network

   This network is characterized by two equations, namely aimed to avoid cycle consistency loss, which is often seen in CycleGAN and StarGAN networks, as well as to avoid identity loss, ensuring little change to the semantic information in the audio data.

The loss functions for these networks are as given below, and explained above, where c from P(c) denotes the categories of emotions. y-p(y—c) and x-P(x) represent the data with attribute c in the training set and the data with any attribute in the training set, respectively.

Hence the final equation for the DEVC model becomes:

$$L_{adv}^{D}(D) = -E_{c\sim p(c),y\sim p(y|c)}[\log D(y,c)] - E_{x\sim p(x),c\sim p(c)}[\log\Big(1 - D\Big(G(x,c),c\Big)\Big)]$$
$$L_{adv}^{G}(G) = -E_{x\sim p(x),c\sim p(c)}[\log D\Big(G(x,c),c\Big)]$$

(a) Generator-Discriminator Network

$$L_{cls}^{C}(C) = -E_{c\sim p(c),y\sim p(y|c)}[\log C(c|y)]$$
$$L_{cls}^{G}(G) = -E_{x\sim p(x),c\sim p(c)}[\log C\Big(c|G(x,c)\Big)]$$

(b) Generator-Classifier Network

$$L_{cyc}(G) = E_{c'\sim p(c),x\sim p(x|c'),c\sim p(c)}\left[\,\|\,G(G(x,c),c') - x\,\|\,_{\rho}\right]$$
$$L_{id}(G) = E_{c'\sim p(c),x\sim p(x|c')}\left[\,\|\,G(x,c') - x\,\|\,_{\rho}\right]$$

(c) Generator-Generator Network

Figure 2.8: Loss Functions for the three adversarial networks

$$L_G = \lambda_{adv}L_{adv}^{G} + \lambda_{cls}L_{cls}^{G} + \lambda_{cyc}L_{cyc}^{G} + \lambda_{id}L_{id}^{G}$$
$$L_D = L_{adv}^{D}$$
$$L_C = L_{clss}^{C}$$

Figure 2.9: Equation for DEVC model

Architectures of DEVC components

Each network is made up of a convolutional residual network, as it can extract deep audio features. The diagram of a C-Block is as follows:

Where Conv2d refers to 2d convolution, and Instance Normalization was selected after observing the benefits of it in image conversion, which is similar to audio conversion. This

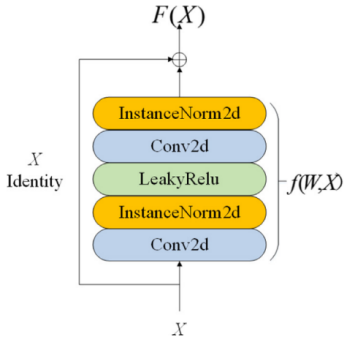
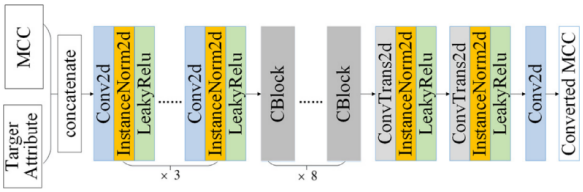


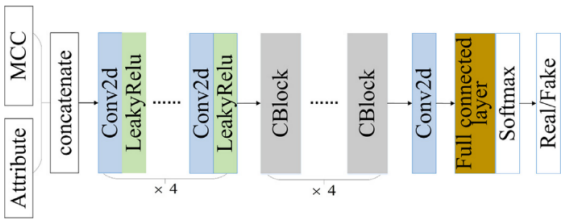
Figure 2.10: C Block Architecture

will lead to better audio quality by normalizing the feature map.

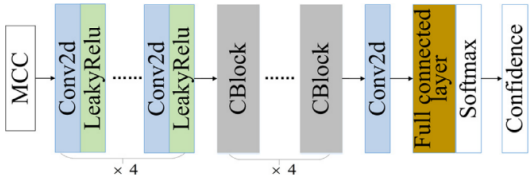
The generator is designed so that it can convert varied lengths of speech, based on convolution residual network. The input and output of each network are as discussed before, and can be illustrated as shown below:



(a) Generator Architecture



(b) Discriminator Architecture



(c) Classifier Architecture

Figure 2.11: CNN Architecture for the three adversarial networks

2.4.3 Speaker Voice Conversion Module

The DSVC is designed to convert the speaker identity of the synthesized speech while retaining the semantic information, such as the content and emotional expression, and is based on a convolutional residual network, which is trained on the VCTK Corpus, a dataset containing speech data from a diverse set of native English speakers. The DSVC is trained to learn the characteristics of different speakers and their unique vocal traits, allowing it to effectively convert the speaker identity of the synthesized speech.

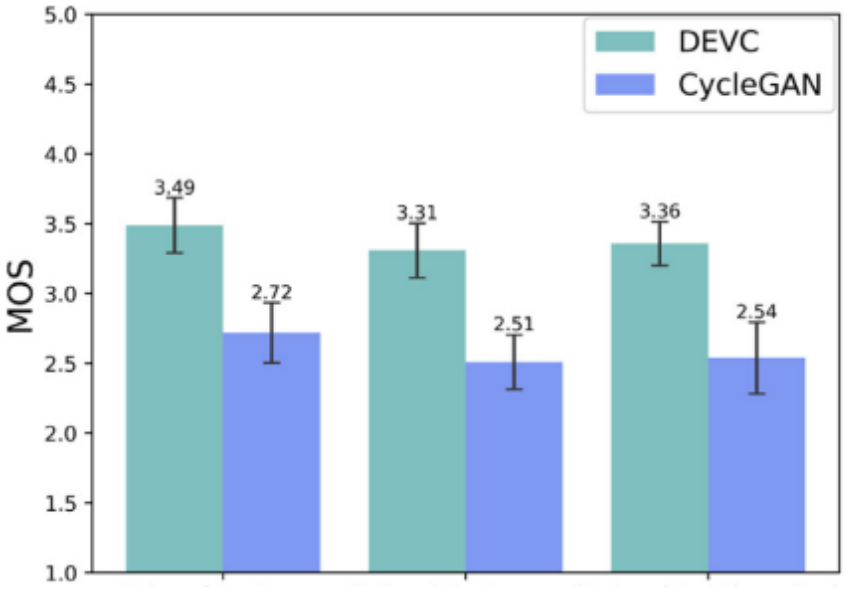
The DSVC has the same structure as the DEVC, albeit it has a deeper network. The generator in the DSVC has 4 extra C-Blocks, while the Discriminator and Classifier have 2 extra C-Blocks. This ensures that the DSVC only extracts and manipulates deep audio features, and does not tamper with the already altered features to bring in the emotion in the DEVC module.

2.4.4 Experimental Analysis

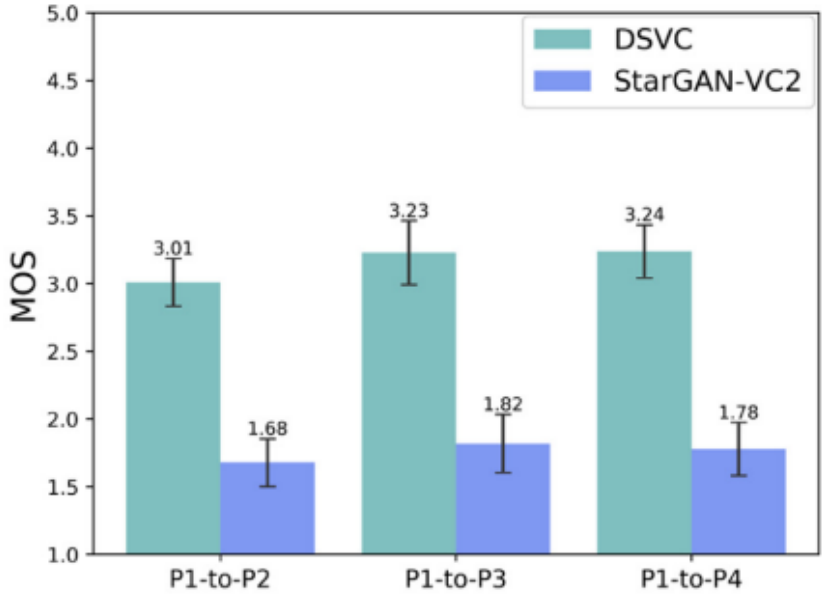
Through quantitative as well as qualitative analysis, it was inferred that the performance of the model was comparable, if not better than the existing models in the area of emotional speech synthesis, as well as speaker identity voice conversion.

Mean Opinion Score Analysis

An analysis of the Mean Opinion Scores(MOS) was conducted to assess the performance of the model for dual-task speech synthesis firstly with emotion voice conversion and secondly, with speaker voice conversion. A higher MOS opinion score denotes that the emotion can be discerned, but is easily not distinguishable as generated speech output or as voiced by humans. It can be seen that the DEVC and DSVC have higher MOS scores than their counterparts, indicating their higher performance levels.



(a) TTS DEVC MOS evaluation



(b) TTS DSVC MOS evaluation

Figure 2.12: MOS scores of DEVC and DSVC models

Mel Spectrum Analysis

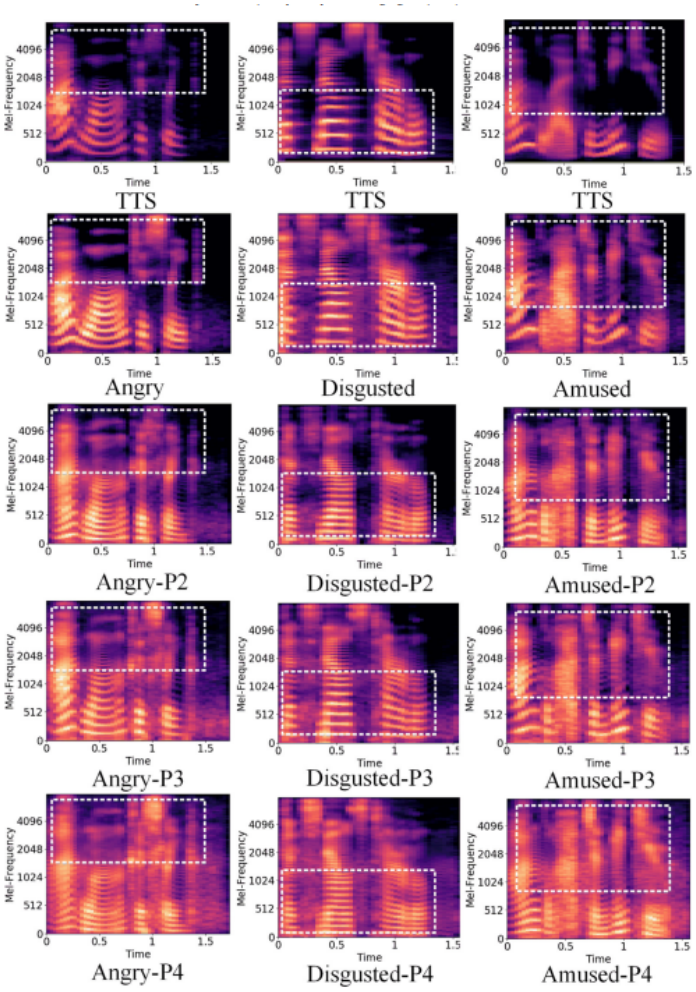


Figure 2.13: Mel Spectrum comparison

On analysis of the Mel Spectrum of various combinations of speaker identities and emotions in this model, several inferences can be made. One of these inferences includes the observation that when converting from neutral to angry, the high energy part of the spectrum becomes increased to denote the louder pitch and the sharpness of the voice. Similarly, it can be observed that the generated disgusted speech spectrum has more energy in the low-frequency part, denoting the deeper pitch that the voice tends to take when disgusted.

2.5 Scaling Speech Technology to 1,000+ Languages

2.5.1 Overview

The PDF titled "Scaling Speech Technology to 1,000+ Languages" presents the Massively Multilingual Speech (MMS) project, which aims to significantly increase the number of languages supported by speech technology. The project focuses on developing pre-trained models, including wav2vec 2.0, automatic speech recognition (ASR), and speech synthesis models, to enable access to speech technology for a vastly expanded range of languages. The MMS project addresses the challenge of scaling speech technology by leveraging labeled datasets, fine-tuning pre-trained models, and developing language identification models for thousands of languages.

2.5.2 Massively Multilingual Speech (MMS)

The paper titled details the Massively Multilingual Speech (MMS) project, which endeavors to significantly expand the number of languages supported by speech technology. The project's primary focus is on developing pre-trained models, including wav2vec 2.0, automatic speech recognition (ASR), and speech synthesis models, to facilitate access to speech technology for a vastly increased number of languages. The MMS project addresses the challenge of scaling speech technology by segmenting data into individual verses, with an average duration of about 12 seconds, and generating posterior probabilities from an acoustic model using a Transformer trained with Connectionist Temporal Classification (CTC). This approach enables the creation of verse-level audio segments for each chapter recording, allowing for efficient forced alignment on GPUs.

The MMS project leverages a labeled dataset to fine-tune pre-trained models for ASR, enabling the transcription of speech in up to 1,107 different languages. Additionally, the project utilizes n-gram language models trained on Common Crawl data using KenLM, with specific adaptations for languages that do not use spaces to separate words. The document also discusses the development of a language identification model for thousands of languages, further contributing to the project's goal of broadening language support in speech technology.The comparison of the MMS project to existing broad coverage approaches and other datasets, such as the CMU Wilderness project and ASR-2K, emphasizing the viability of the new data for building machine learning models. The MMS

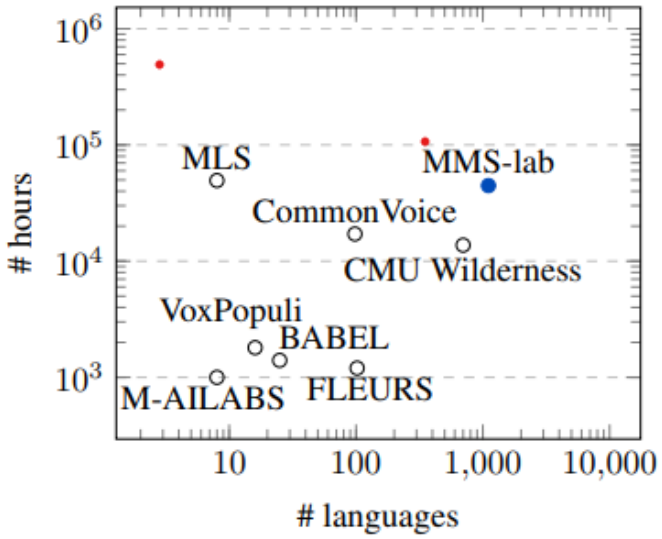


Figure 2.14: Labelled Datasets

project's forced alignment process is distinguished by its global alignment approach, which considers the entire sequence, in contrast to segment-based alignments. The document also references related studies, including "Attention is all you need" and "VoxPopuli," which contribute to the broader context of multilingual speech corpus development and representation learning.

The MMS project represents a significant advancement in scaling speech technology to a vastly expanded range of languages, with a focus on efficient forced alignment, pre-trained models for ASR and speech synthesis, language identification, and the development of language models tailored to diverse linguistic structures. The project's findings and results demonstrate promising progress in reducing word error rates and enhancing access to speech technology for a more extensive array of languages, thereby contributing to increased inclusivity and accessibility on a global scale.

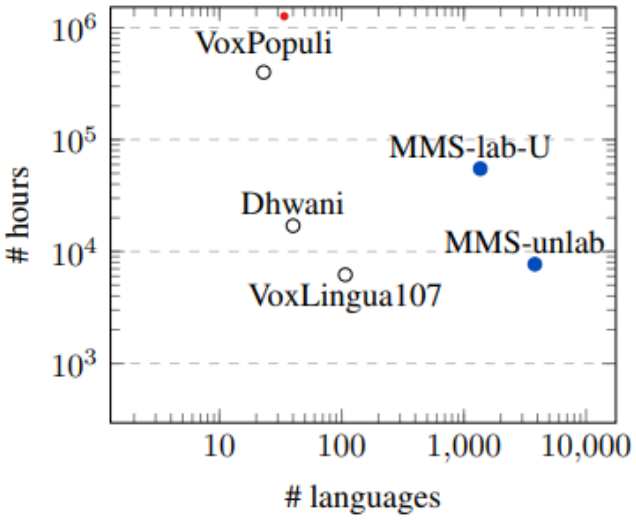


Figure 2.15: Unlabelled Datasets

2.5.3 Comparison to Existing Broad Coverage Approaches and Other Datasets

The CMU Wilderness project [Black, 2019] also used New Testament data to build speech synthesis models for 699 languages and ASR-2K focused on automatic speech recognition for nearly two thousand languages. Finally, we assess the viability of the new data for building machine learning models by comparing the performance of ASR models trained on MMS-lab to models trained on an existing dataset in an out-of-domain setting.

To compare the effectiveness of their data creation method, we take the original data from the data source and apply either our protocol or the protocol of Black. For languages where multiple recordings exist, we only use the recordings used in the CMU Wilderness dataset to enable a better comparison. Next, we use the resulting data to fine-tune XLS-R models for monolingual ASR and then measure accuracy in terms of character error rate on the FLEURS dev set.

The figure 2.16 shows MMS-lab data preparation process results in better quality ASR models compared to CMU Wilderness with improvements between 2.1 percentage - 4.7 percentage CER, depending on the language. Our alignment procedure also retains a

much larger amount of the training data compared to the CMU Wilderness protocol: for Telugu, there are 26.5 hours of data, MMS-lab retains 26.2 hours compared to 11.1 hours for the CMU Wilderness process. For English, we start with 17.3 hours, MMS-lab retains 17 hours vs. 10.6 hours for CMU Wilderness

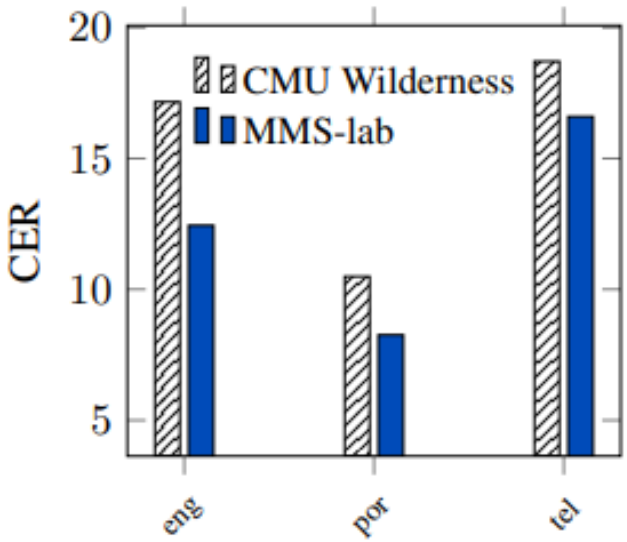


Figure 2.16: MMS-lab vs. CMU Wilderness. Character Error Rate of ASR models in English (eng), Portuguese (por) and Telugu (tel) on the FLEURS dev set.

2.6 Summary and Gaps Identified

2.6.1 Summary

Paper Title	Advantages	Disadvantages
An efficient adaptive artificial neural network based text to speech synthesizer for Hindi language	1. High prediction accuracy. 2. Utilizes linguistic constraints. 3.Potential applications in accessibility, language learning, assistive communication, multimedia, and research and development.	1. Lack of validation and generalization of the proposed approach 2. Limited discussion of ethical considerations related to the development and use of the text to speech synthesizer 3. Limited benchmarking against existing speech synthesizers for the Hindi language
Classification of emotions from speech signals using machine learning	1.Provides a broad overview of AI/ML applications in citizen science. 2.Discusses potential benefits and risks. Advocates for technology investments in the environment.	1. Generalized discussion, lacks in-depth exploration of specific projects 2.Limited discussion on ethical considerations or potential downsides.

Paper Title	**Advantages**	**Disadvantages**
An efficient sentimental analysis using hybrid deep learning and optimization technique for Twitter using parts of speech (POS) tagging	1. The proposed hybrid deep learning approach achieves better accuracy and efficiency compared to other commonly used models, as indicated by the 92 percentage accuracy and 91.3 percentage precision achieved by the artificial neural network (ANN) 2.The article provides a comprehensive review of related works in sentiment analysis, providing valuable insights into the existing research landscape and the state-of-the art techniques.	1.The article lacks a discussion on potential limitations or challenges of the proposed hybrid deep learning approach, which could provide a more balanced view of the technique 2.The article does not explicitly address the generalizability of the proposed technique to sentiment analysis in languages other than English, which could limit the applicability of the approach in multilingual contexts
MASS: Multi-task emotion and speaker conversion	1.Can be used for speaker and emotion conversion	1. he MMS-lab dataset, while covering a large number of High training cost.

Paper Title	**Advantages**	**Disadvantages**
Scaling Speech Technology to 1,000+ Languages	The MMS significantly expands language coverage in speech technology, supporting up to 1,107 languages for automatic speech recognition (ASR) and speech synthesis, as well as language identification for 4,017 languages. 2. Experiments show that the multilingual speech recognition model developed by the MMS project more than halves the word error rate of existing benchmarks on 54 languages, despite being trained on a small fraction of labeled data. 3. The development of pretrained wav2vec 2.0 models covering 1,406 languages and a single multilingual ASR model for 1,107 languages provides a valuable resource for researchers and developers working on multilingual speech technology.	1. he MMS-lab dataset, while covering a large number of languages, is primarily sourced from a specific narrow domain, potentially limiting the performance of the models when applied to other domains or unseen speakers. 2. Most recordings in the MMS-lab dataset are from a single speaker, which may impact the models' ability to generalize to diverse speech patterns and accents, particularly in languages with significant regional variations. 3. Training machine learning models on religious texts, as done in the MMS project, may introduce biases, requiring careful ethical considerations and analysis to ensure fair and unbiased performance across different genders and religious contexts.

Table 2.3: Summary of scientific papers

2.6.2 Gaps Identified

1. Generalizability: While the proposed hybrid deep learning (HDL) approach in sentiment analysis on Twitter demonstrates impressive performance, the research does not explicitly address potential challenges or limitations related to the cultural and contextual nuances in sentiment expression on social media, which could impact the generalizability of the model across diverse user groups or topics.

2. Lack of Specifics on Emotion Analysis Techniques: The text mentions the use of mel-spectrums to analyze emotions but lacks specific details on the techniques employed for emotion analysis. This lack of information may lead to questions about the robustness and accuracy of the emotion analysis module.

3. Dependency on Speech Quality: The effectiveness of the proposed method could be influenced by the quality of the speech signals. It would be valuable to address how the method performs with variations in speech quality, such as low-quality recordings or distorted signals.

4. Sensitivity to Linguistic Variations: The text mentions the use of linguistic restrictions and features from Mel-frequency cepstral coefficients (MFCC) to model factors like intonation, duration, and syllable intensities. However, the model's performance may be sensitive to linguistic variations within the Hindi language, including regional accents or dialects. The study does not explicitly address the impact of such variations.

Chapter 3

Requirements

3.1 Hardware and Software Requirements

3.1.1 Hardware

- Multi-core processor with a clock speed of at least 2.5 GHz or higher
- 8GB of RAM or more
- Hard Disk as Storage (preferably SDD)
- Speaker for audio output
- Network connecting device

3.1.2 Software

- Python: PyTorch, TensorFlow, Keras, iNLTK, Librosa
- HTML5, CSS
- XAMPP

3.2 Functional Requirements

1. **User interface:** The Emotive Malayalam Text-to-Speech Synthesizer has a user-friendly interface that allows users to input text, adjust speech parameters, and preview synthesized audio.
2. **Authentication:** The authentication process for the Emotive Malayalam Text-to-Speech Synthesizer ensures secure access to the system, protecting user data and maintaining the integrity of the application.

3. **Input text:** The input text for the Emotive Malayalam Text-to-Speech Synthesizer refers to the written content or script that users input into the system for conversion into emotive speech. This text serves as the foundation for generating expressive audio output with specific emotional characteristics.

4. **Emotion Identification:** The emotion identification module in the Emotive Malayalam Text-to-Speech Synthesizer is responsible for analyzing the input text to identify the underlying emotions present within the content. This module has two models, the first being a translation model that converts the Malayalam inuot text into its English equivalent. The second model utilizes emotion analysis models trained with specific databases to accurately recognize and categorize emotions such as happiness, sadness, anger, and more within the provided text.

5. **Text to Speech:** The text-to-speech module in the Emotive Malayalam Text-to-Speech Synthesizer is designed to convert the input text into neutral speech using a TTS model. This process involves analyzing the text and generating synthesized speech with a neutral tone, serving as the basis for the subsequent emotion voice conversion module.

6. **Emotion Voice Conversion:** The emotion voice conversion module in the Emotive Malayalam Text-to-Speech Synthesizer is responsible for synthesizing emotion in speech using the identified emotion from the analysis and generated speech from the text-to-speech module. This module utilizes emotion voice conversion models trained with specific databases in high resource languages, to modify the neutral speech generated by the text-to-speech module to convey the desired emotional characteristics. The resulting audio output is expressive and emotive, providing a valuable tool for individuals with specific communication needs and enhancing emotional impact in audio content

7. **Error Handling:** Implement robust error handling for various scenarios, such as incorrect file formats, server errors, or network issues.

8. **Security:** The Administrator defines and configures security policies for the software application.

Chapter 4

System Architecture

This chapter explores the framework and structural design that constitute the basis of the project.

4.1 System Overview

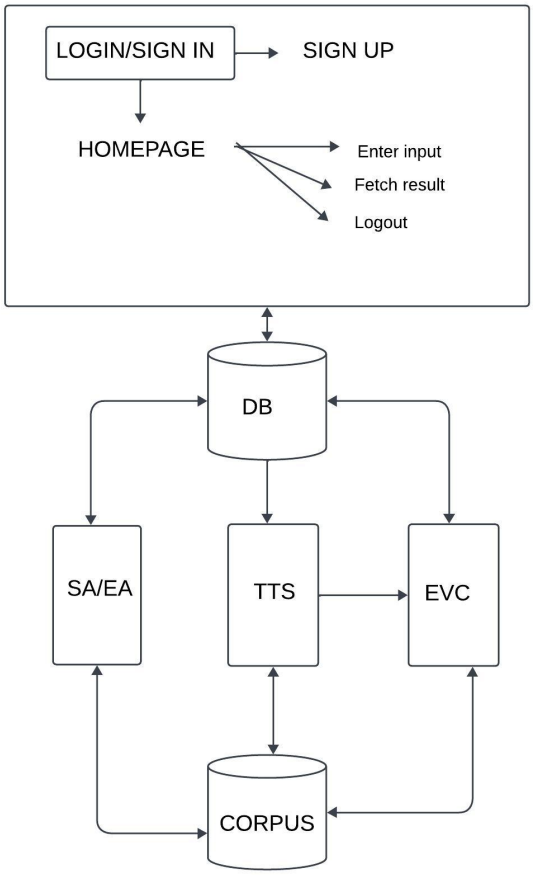


Figure 4.1: Architecture Diagram

1. **Emotional analysis module**

The model is used to learn the relationship between the features extracted from text data and the probability of each emotion class. Let's denote $\mathbf{x}$ as the input feature vector representing the text data, and y as the emotion class label. The logistic regression model estimates the conditional probability $P(y|\mathbf{x})$ using a logistic function.

The logistic function, also known as the sigmoid function, is defined as:

$$\sigma(z) = \frac{1}{1 + e^{-z}}$$

Where $z = w_0 + w_1x_1 + w_2x_2 + \ldots + w_nx_n$ is the linear combination of input features weighted by coefficients $\mathbf{w}$, and n is the number of features.

The logistic function squashes the output of the linear combination to the range [0, 1], representing probabilities. In the context of emotion classification, each class is associated with its set of weights $\mathbf{w}$, and the model calculates the probability of each class given the input text features.

The model is trained by optimizing the weights $\mathbf{w}$ to minimize the cross-entropy loss function, which measures the dissimilarity between the predicted probabilities and the actual class labels. This is done using optimization algorithms like gradient descent.

During inference, the model predicts the emotion class with the highest probability among all classes for a given input text. logistic regression for emotion classification learns a linear relationship between text features and the probability of each emotion class using the logistic function, and it's trained to minimize the discrepancy between predicted and actual class probabilities.

2. **Text-to-speech synthesis**

Text-to-Speech (TTS) synthesis is a crucial component in various applications, enabling the conversion of written text into spoken words. This report provides an overview of the TTS module, outlining its key components and underlying equations.

$$F(T) = TextAnalysis(T) \tag{4.1}$$

Prosody Generation

$$G(P) = ProsodyGeneration(P) \tag{4.2}$$

Speech Synthesis

$$S = SpeechSynthesis(F(T), G(P)) \tag{4.3}$$

Equations for Mel-frequency Cepstral Coefficients (MFCC)

In many TTS systems, Mel-frequency cepstral coefficients (MFCC) are employed for feature representation. The MFCCs are computed using the following steps:

- Frame the speech signal into overlapping frames.
- Apply a window function to each frame.
- Compute the Fast Fourier Transform (FFT) of each framed signal.
- Apply Mel filterbanks to the FFT results.
- Take the logarithm of the filterbank energies.
- Apply the Discrete Cosine Transform (DCT) to obtain MFCCs.

The MFCCs are given by:

$$MFCC(n) = \sqrt{\frac{2}{N}} \sum_{m=1}^{M} \log(Energy(m)) \cos\left(\frac{\pi n(2m-1)}{2N}\right) \tag{4.4}$$

where N is the number of MFCC coefficients, M is the number of filterbanks, and $Energy(m)$ represents the energy in the m-th Mel filterbank.

3. **Emotion voice conversion**

The emotion voice conversion process can be mathematically represented using a mapping function. Let f_{EVC} be the Emotion Voice Conversion function, mapping source features X to target features Y with respect to a particular emotion:

$$Y = f_{EVC}(X, Emotion) \tag{4.5}$$

Here, *Emotion* represents the target emotion to be infused into the synthetic speech.

Feature Extraction

To achieve emotion voice conversion, relevant features need to be extracted from the source and target voices. Commonly used features include Mel-frequency cepstral coefficients (MFCC), pitch, and energy. The MFCC features, denoted as X_{MFCC} and Y_{MFCC} for the source and target voices respectively, are particularly crucial in capturing spectral characteristics.

$$X_{MFCC} = MFCC(X) \tag{4.6}$$

Conversion Process

During the conversion process, the trained model is applied to convert the source features to target features based on the specified emotion.

$$Y_{MFCC\_converted} = f_{EVC}(X_{MFCC}, \theta, Emotion) \tag{4.7}$$

Synthetic Speech Synthesis

The final step involves synthesizing speech using the converted features. This can be achieved using a vocoder or waveform generation algorithm.

$$SyntheticSpeech = Vocoder(Y_{MFCC\_converted}) \tag{4.8}$$

Evaluation Metrics

The performance of the Emotion Voice Conversion module can be evaluated using objective metrics such as Mean Opinion Score (MOS), root mean square error (RMSE), and mel-cepstral distortion (MCD).

$$MOS = \frac{1}{N}\sum_{i=1}^{N} Rate_i \tag{4.9}$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(Y_i - \hat{Y}_i)^2}{N}} \tag{4.10}$$

$$MCD = \frac{10}{\ln(10)}\sqrt{2\sum_{i=1}^{N}(MFCC_i - M\hat{F}CC_i)^2} \tag{4.11}$$

Here, N represents the number of samples, Y_i is the true value, $\hat{Y}_i$ is the predicted value, $MFCC_i$ is the original MFCC value, and $M\hat{F}CC_i$ is the converted MFCC value.

The effectiveness of the Emotion Voice Conversion module is determined by the ability to achieve a high MOS, low RMSE, and low MCD values.

4.2 Architectural Design

4.2.1 Sequence Diagram

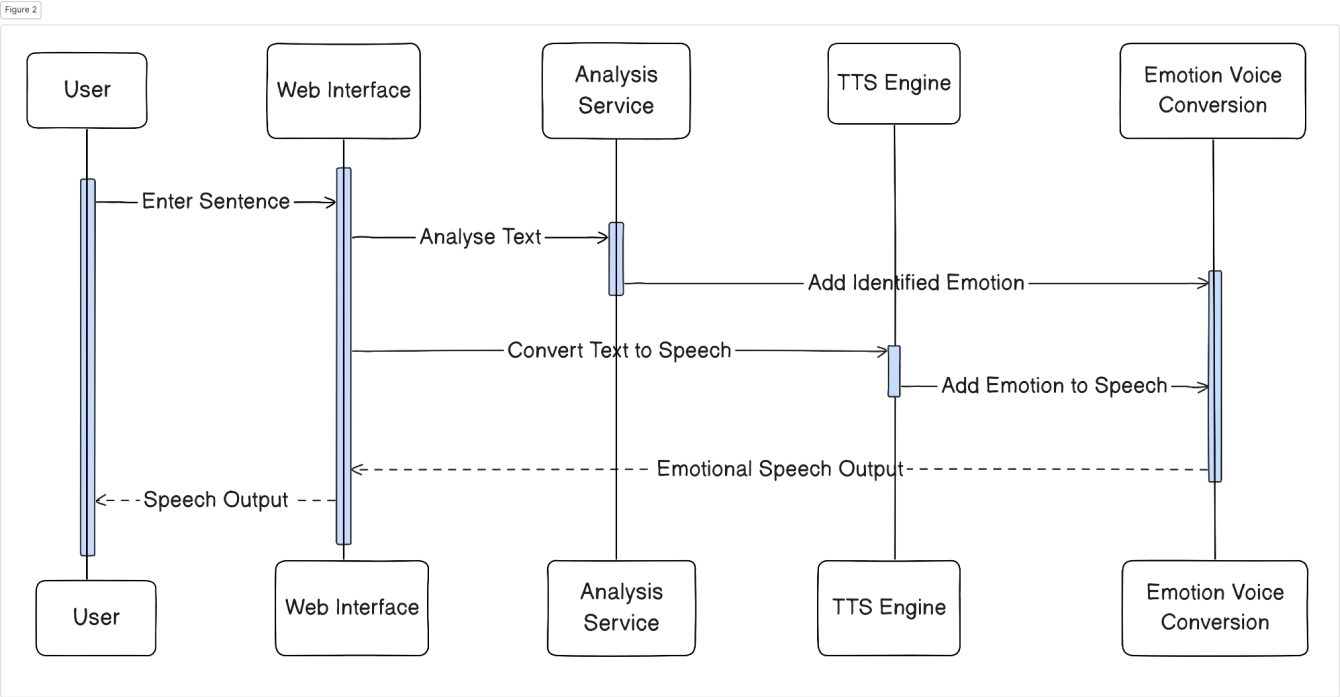


Figure 4.2: Sequence Diagram

4.3 Module Division

1. **Emotion Analysis:**

The proposed system for emotion detection in Malayalam text involves two sub-modules. Translation module and Emotional analysis module.

i. Translation Module

This module utilizes the Google Translator API(Application Program Interface) version 4.0.0-rc1 for translating the Malayalam text as input from the user to English text. Output of this module is used as the input for the next emotional analysis module.

ii. Emotional Analysis Module

This module is a Logistic Regression model trained using data corpus made with different emotion classes for each sentence. Initially, the data is preprocessed by removing stopwords, punctuations, hashtags, user handles, etc and it is splited into training and testing data. The data is then tokenized and fed into the multinomial LogisticRegression model for training. The model attain a 90% accuracy for classifying emotions, signifying its proficiency in correctly assigning emotion classes to text inputs. This emotion identified is stored and later used in the emotion voice conversion module.

2. **Text to Speech Synthesis**

The converts written text into spoken language, utilizing linguistic analysis and voice synthesis techniques. It plays a pivotal role in applications ranging from accessibility tools for the visually impaired to enhancing natural human-computer interaction.

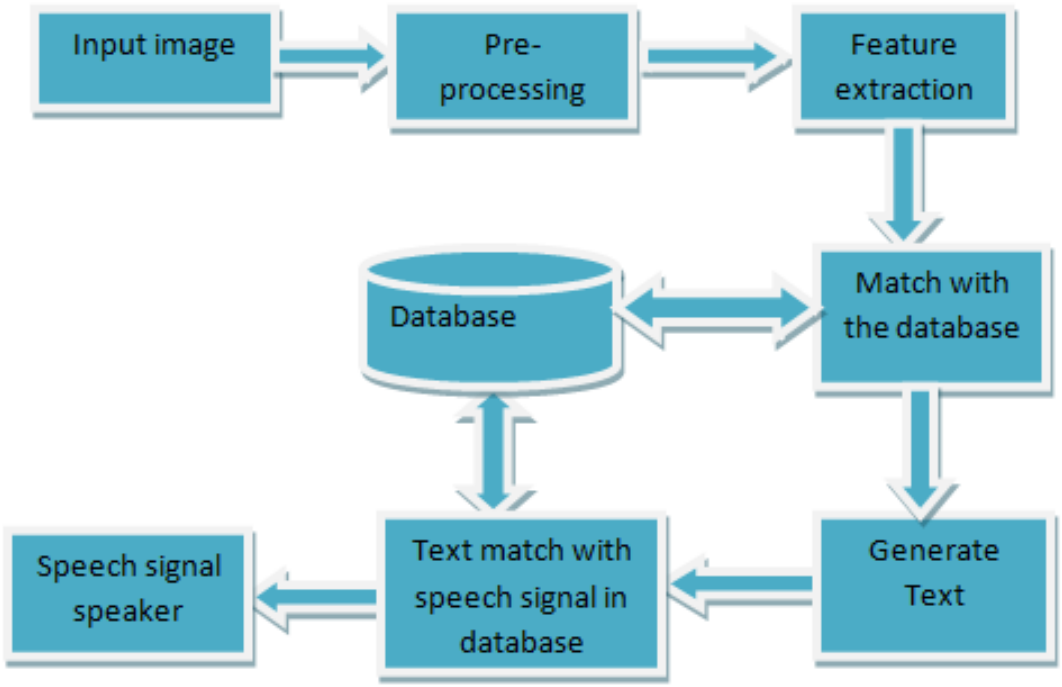


Figure 4.3: Text to Speech Synthesis

3. Emotion Voice Conversion

The Emotion Voice Conversion module in Text-to-Speech (TTS) transforms the neutral voice into one with specified emotions, enhancing the expressiveness of synthetic speech. Utilizing algorithms and statistical models, it maps source voice features to emotional target features, contributing to a more emotionally resonant and personalized TTS experience. This model is based on the STAR-GAN VC2 model that performs many-to-many voice conversion ideally for speaker voice conversion. However, the modified model was trained in order to perform emotion voice conversion.

The dataset used for training the model was the Emotional Speech Dataset, with 3500 audio samples and 10 different speakers, in both Mandarin and English. The structure of the model is detailed in the figure below.

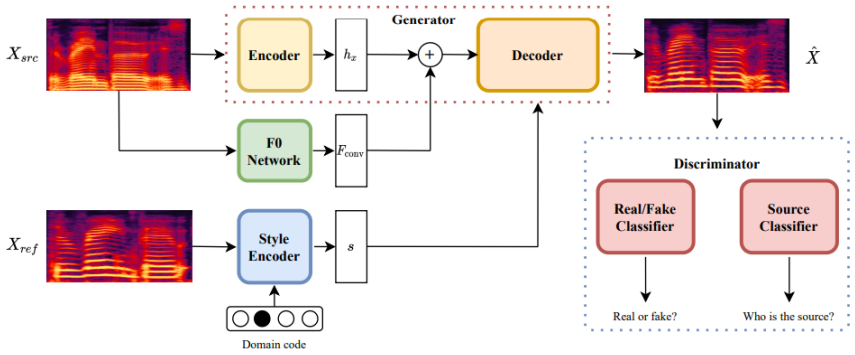


Figure 4.4: Emotion Voice Conversion

In the figure 4.4, Xsrc is the source input, Xref contains the style information and acts as a reference input. The latent feature, and fundamental frequency of the source are concatenated by the channel as input into the decoder.There are two classifiers that aim to distinguish between real and fake samples, depending upon the mel-spectogram of the source input.

The model consists of 5 different components. The generator converts and input mel-spectogram into a modified mel-spectogram that mimics the style, in this case emotion, and produces a new sample according to the reference provided. The F0 extraction network exists to obtain the fundamental frequency(F0) of any given voice sample. There is also a mapping network, that maps a given style with a Gaussian randomized latent code, so that there is a diverse representation of all the domains in the dataset. The style encoder has a similar function to that of the mapping network, except that its purpose is to extract a style code from a given domain. Lastly, the discriminator's function is to distinguish between a real and generated sample.

The goal of the model is to find the mapping between the source domain which is always neutral in the scope of the project, to the target domain, or emotion classes. The model is trained based on the following loss functions.

(a) Adversarial Loss:

$$\begin{aligned}\mathcal{L}_{adv} = \; & \mathbb{E}_{\boldsymbol{X},y_{src}}\left[\log D(\boldsymbol{X}, y_{src})\right] + \\ & \mathbb{E}_{\boldsymbol{X},y_{trg},s}\left[\log\left(1 - D(G(\boldsymbol{X}, s), y_{trg})\right)\right]\end{aligned}$$

The adversarial loss is what the generator learns from in order to produce more realistic samples. Here, D denotes the domain, where X is the input, s the style and Y the source and target domains according to the subscript.

(b) Cycle Consistency Loss:

$$\begin{aligned}\mathcal{L}_{adv} = \; & \mathbb{E}_{\boldsymbol{X},y_{src}}\left[\log D(\boldsymbol{X}, y_{src})\right] + \\ & \mathbb{E}_{\boldsymbol{X},y_{trg},s}\left[\log\left(1 - D(G(\boldsymbol{X}, s), y_{trg})\right)\right]\end{aligned}$$

The cycle consistency loss is what the generator learns from in order to preserve all the other features of the source sample, such as the content of the audio.

(c) Style Reconstruction Loss:

$$\mathcal{L}_{sty} = \mathbb{E}_{\boldsymbol{X}, y_{trg}, s}\left[\|s - S(G(\boldsymbol{X}, s), y_{trg})\|_1\right]$$

The style reconstruction loss is used so that the style can be reconstructed from the sample that is produced by the generator.

4.3.1 Work Breakdown and Responsibilities

Member Name	Assigned Module
Akash Vijay	Emotion Analysis, Text-to-Speech Synthesis, Back-end
Aleena Karatra	Emotion Voice Conversion, Front-end UI
Alvin George Viji	Text-to-Speech Synthesis, Emotion Analysis, Front end
Ashly Sabu	Emotion Voice Conversion, Emotion Analysis, Back-end

Table 4.1: Assignment of Modules to Team Members

4.4 Work Schedule- Gantt Chart

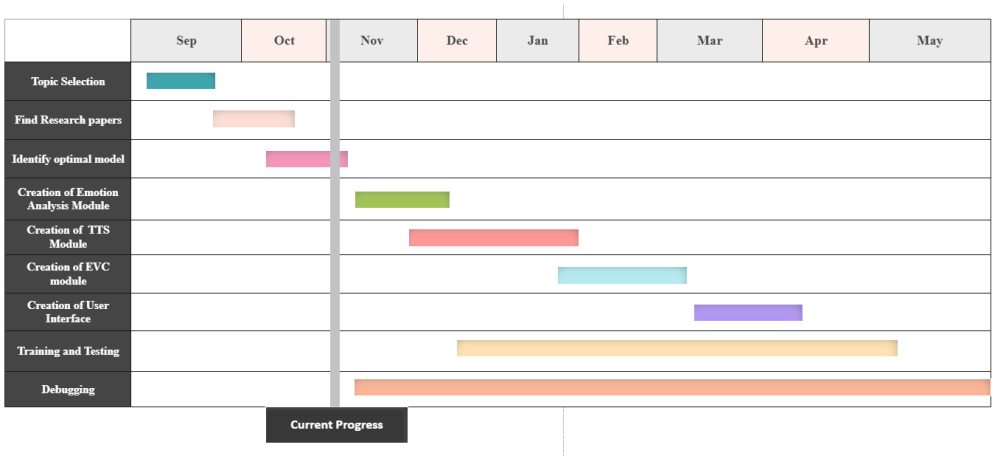


Figure 4.5: Gantt Chart

Chapter 5

Result and Discussion

5.1 Overview

Through the implementation of this project, the Emotional Malayalam Text-to-Speech Synthesizer, or Vikara, several conclusive results were obtained. As a standalone project for the language of Malayalam, there is no other project that this particular project can be compared with in order to compare its results. However, the obtained results were analysed with the help of careful graphical analysis of the produced waveforms, and the comparison of the generated waveforms with the initial waveforms.

```
text = "മലയാളം ഭാഷ ചരിത്രം."
inputs = tokenizer(text, return_tensors="pt")
```

```
from IPython.display import Audio

Audio(output, rate=model.config.sampling_rate)
```

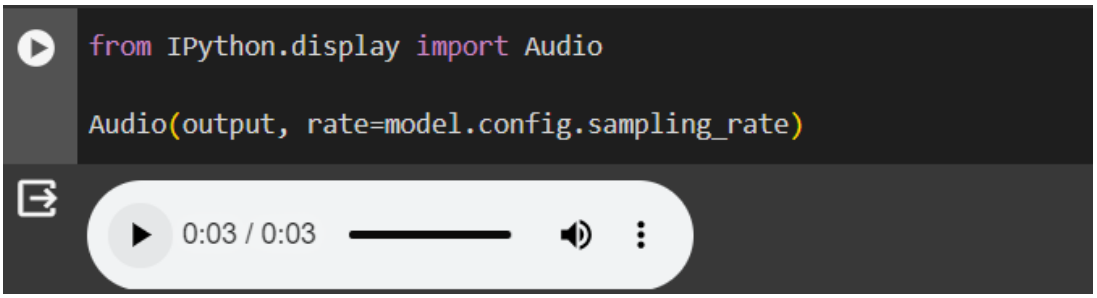


Figure 5.1: Text to Speech

```
emotion_prediction("ഇന്ന് ഒരു മനോഹരമായ ദിവസമായിരുന്നു എന്നാൽ ഉച്ചയോടെ അത് മോശമായി തുടങ്ങി ")
Prediction: joy, Prediction Score: 0.6257849335670471
```

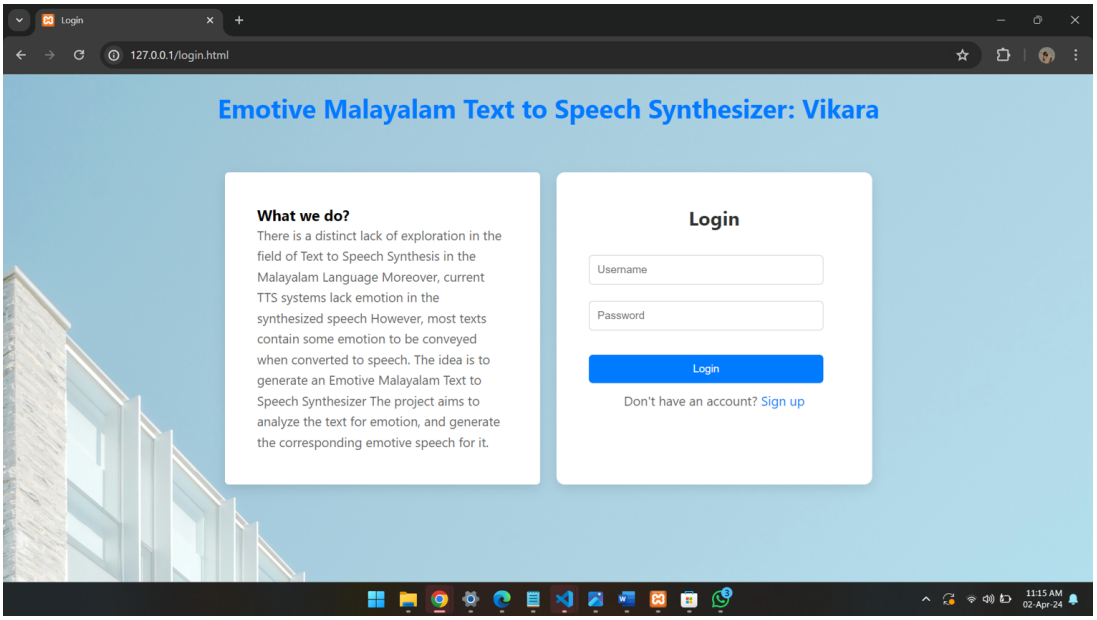


Figure 5.2: GUI1

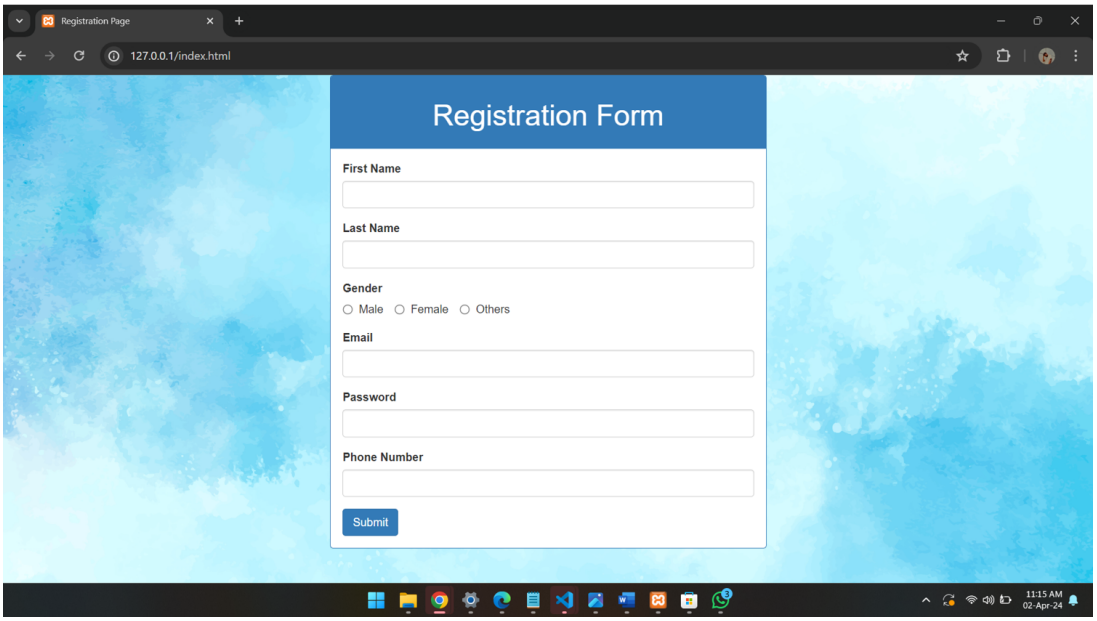


Figure 5.3: GUI2

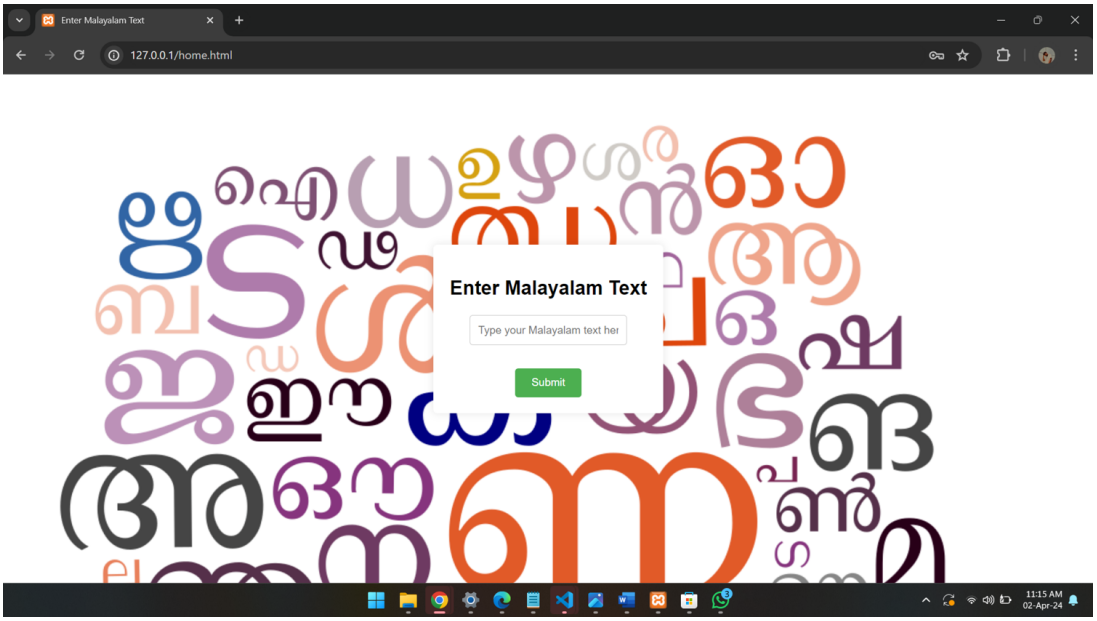


Figure 5.4: GUI3

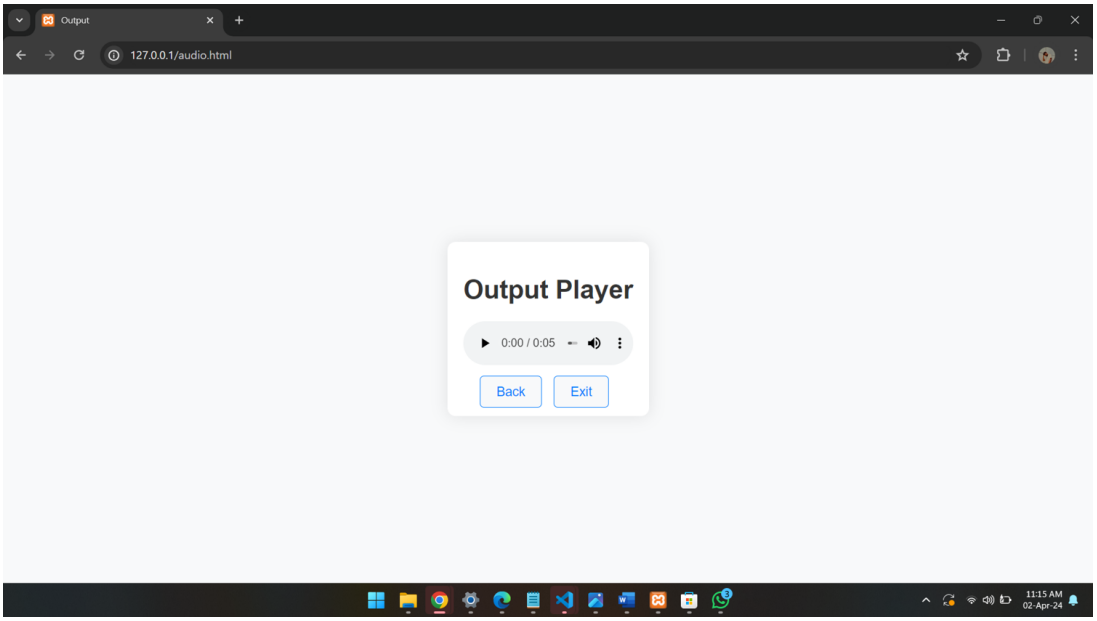


Figure 5.5: GUI4

5.2 Quantitative Results

A specific quantitative result cannot be obtained for both the emotion identification model as well as the emotional voice conversion model, as both models do not directly work with a dataset in the Malayalam language, and are instead trained on a high resource language including those of English and Chinese.

The emotional analysis model is trained using the Logistic Regression model using five classes of emotional data in English. The logistic regression model was able to predict the sentiment with 89.1 percent accuracy. The Malayalam text input was translated into English text using the translation API and was able to accurately identify its sentiment.

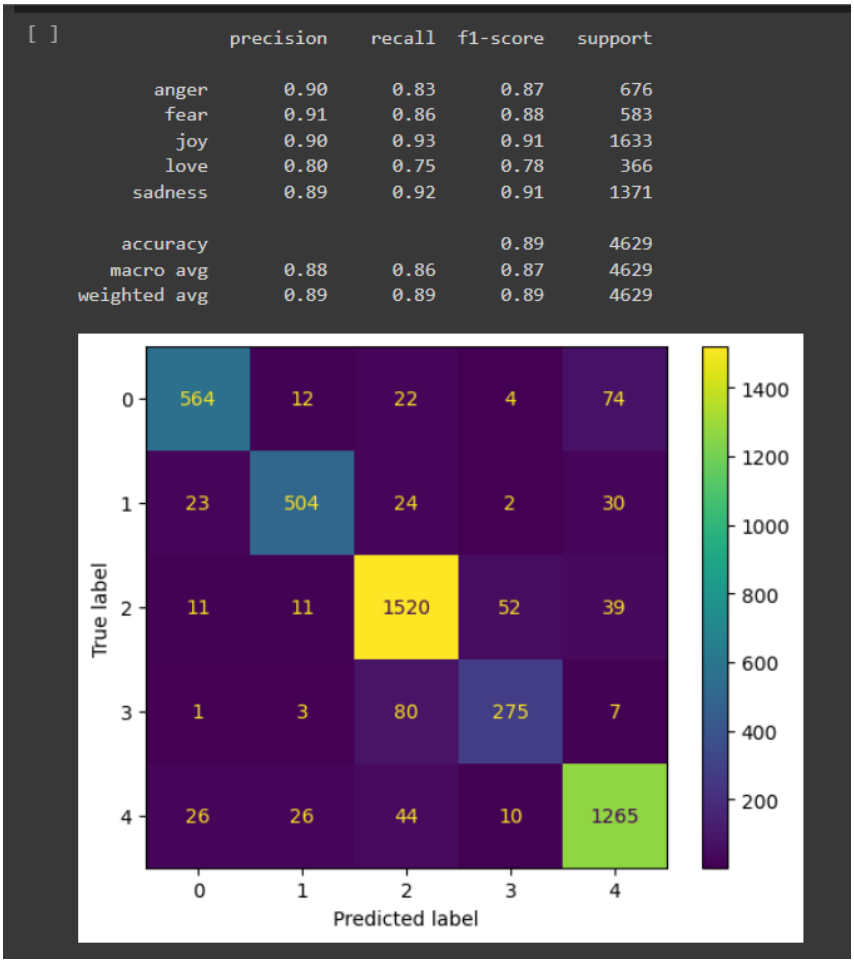


Figure 5.6: Emotion Analysis- Quantitative Results

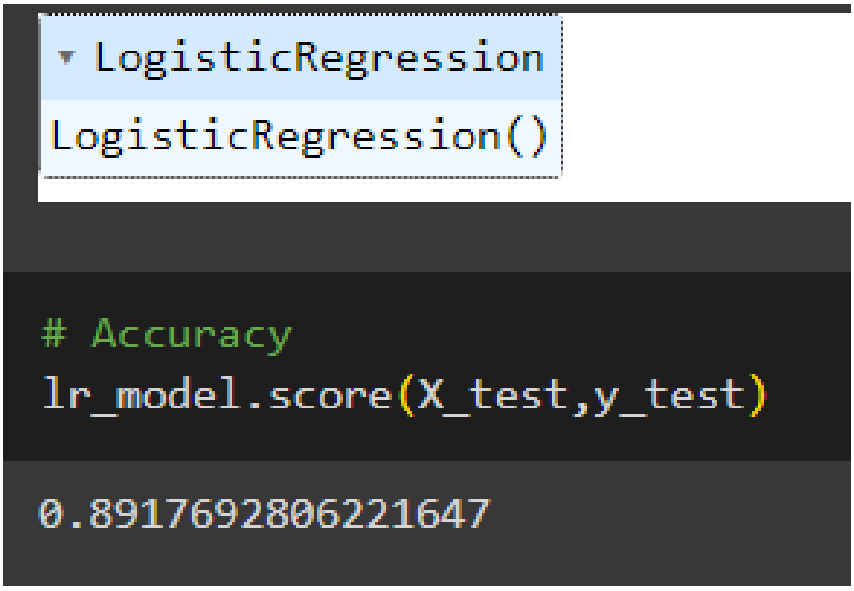


Figure 5.7: Emotion Analysis- Quantitative Results

5.3 Graphical Analysis

Figure 5.6 explains the output score obtained after predicting the emotion from the input text. The model successfully identified different emotions in the text along with its prediction score and it is displayed within a bar plot graph. It can be observed that the emotion 'Joy' has the highest value of the predicted score among the rest of the emotion scores.

And from the other diagrams below, it can be inferred that the emotion voice conversion model is able to correctly convert the input audio into the required emotional audio form. The graphs in the figure 5.7 and 5.8 clearly show the difference in the waveforms of a neutral speech, synthesized by a normal TTS, and the converted speech. In this example, the speech is converted from neutral to angry. Changes can be seen in the output waveform, that correspond to the emotion class 'Anger' as the amplitude and frequency characteristics of the audio are modified to reflect that of an angry person. The amplitude increases, and the frequency increases as well.

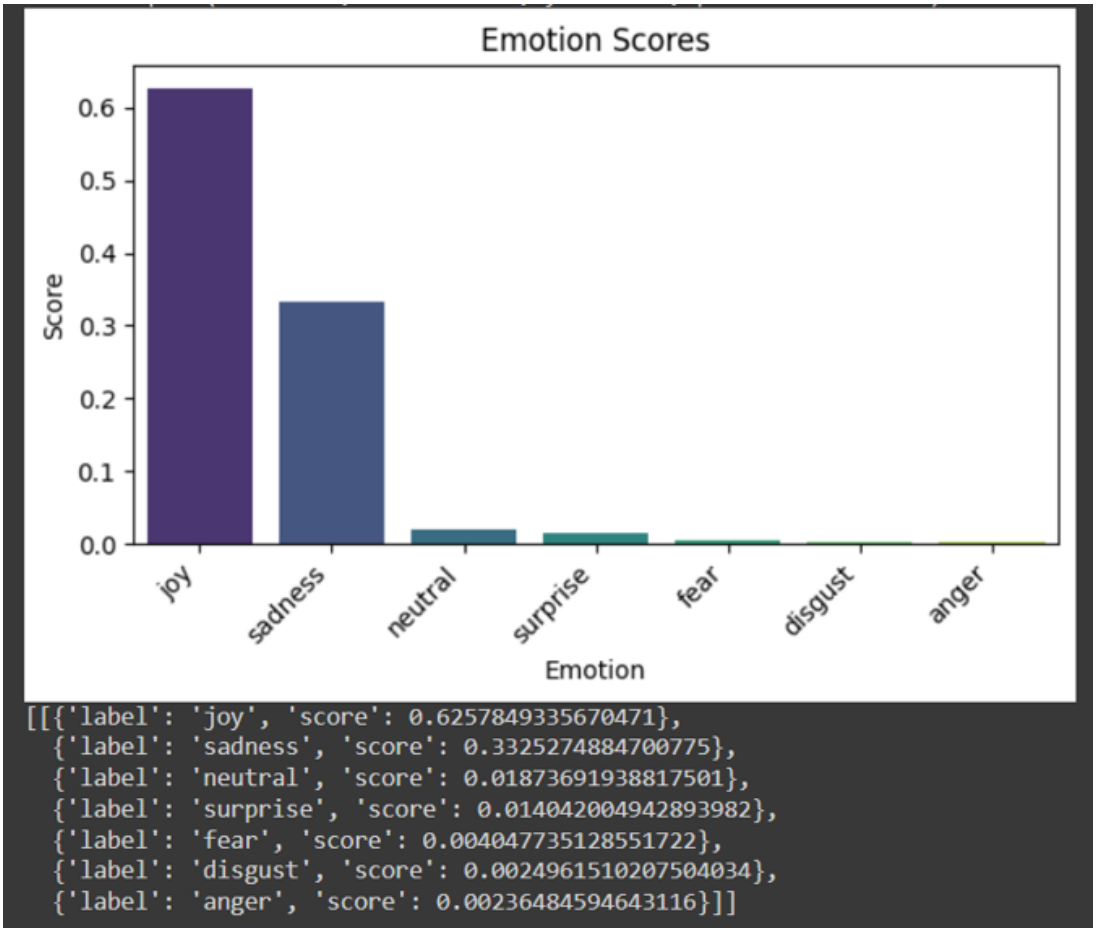


Figure 5.8: Emotion Analysis

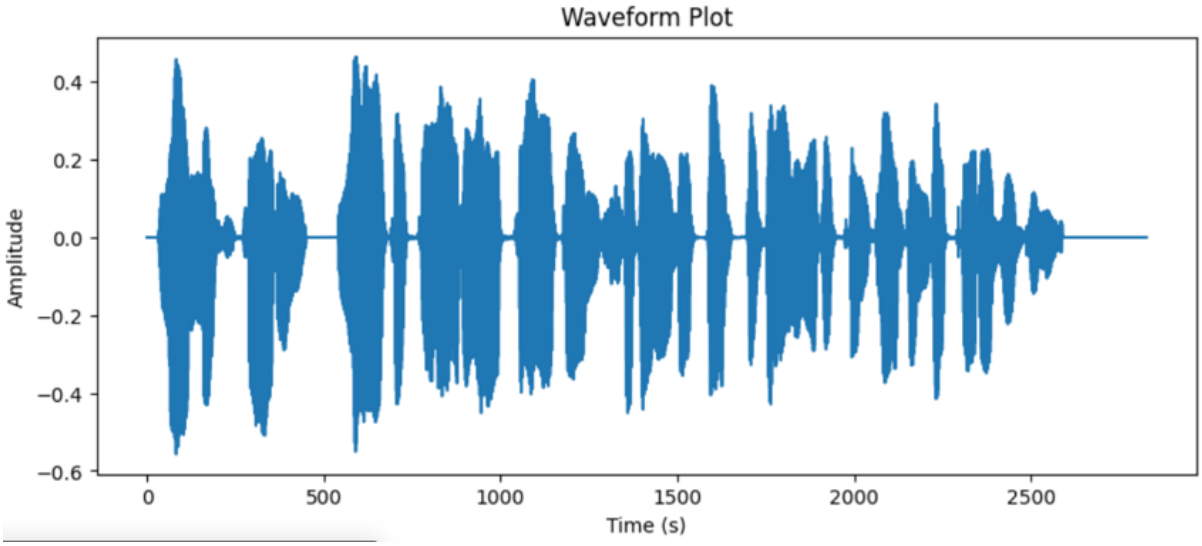


Figure 5.9: Emotion Voice Conversion: Input

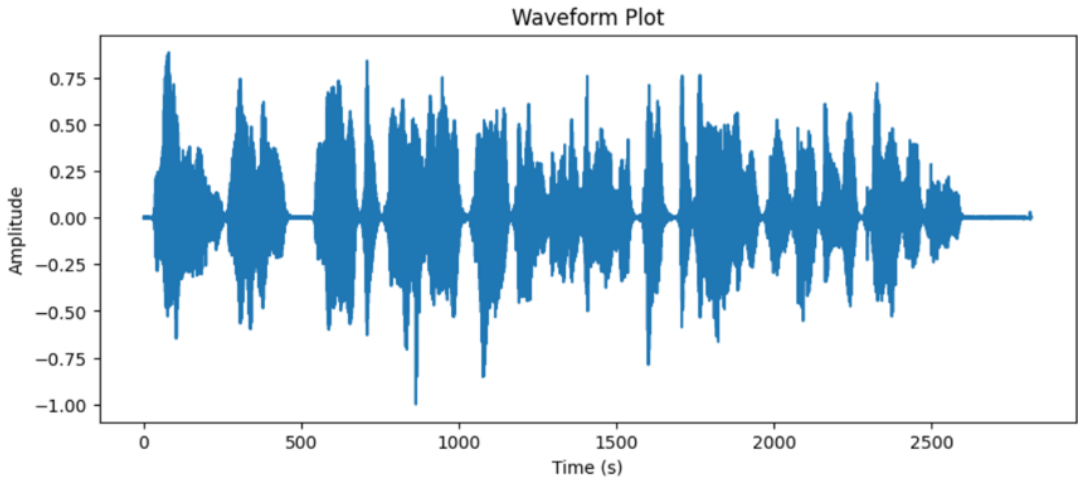


Figure 5.10: Emotion Voice Conversion: Output

5.4 Discussion

The results discussed in this chapter provide conclusive evidence that the project was able to effectively synthesize emotional Malayalam speech from a sample Malayalam text. As a unique venture, the project does not have any specific models of comparison. However, result analysis was made possible through graphical analysis. The initial phases of the project expected more conclusive evidence, however, the lack of corpus availability for the language of Malayalam posed a significant challenge to the completion of the project, causing slight deviations from the expected results. Hence, the use of translation models, and not directly training with a Malayalam dataset has altered the expected results form the project. Creation of a emotional text and speech corpus in the Malayalam language remains as a future scope for improvement.

Chapter 6

Conclusion and Future Scope

6.1 Conclusion

The development of an emotive Malayalam Text-to-Speech (TTS) synthesis system entails multiple crucial steps to infuse emotional expression, depth, and nuance into the generated speech. This process involves translating Malayalam text into English text, annotating emotional content, and incorporating emotion markup within the text. The TTS model is trained to understand and reproduce the prosody associated with different emotions, utilizing deep learning architectures like recurrent neural networks or transformers. Fine-tuning the model using Malayalam emotional speech data is essential to ensure accurate and natural emotional expression. Once developed, this emotive TTS system finds applications across various domains. It can be integrated into audio content creation tools to enhance the emotional impact of podcasts and audiobooks. Furthermore, the system can be employed in conversational agents, such as chatbots and virtual assistants, to create more engaging and human-like interactions. It also serves as a valuable tool for individuals with specific communication needs, offering customization options to adapt emotional expression. Additionally, educational platforms, language learning applications, and entertainment domains can benefit from the system to provide a more immersive and emotionally resonant experience. In customer service and interactive systems, the incorporation of emotional TTS enhances automated interactions, making communication more empathetic and user-friendly. Overall, the emotive Malayalam TTS system enriches a wide array of applications by adding a nuanced and emotionally resonant dimension to synthesized speech.

6.2 Future Scope

The future scope for the Emotive Malayalam Text-to-Speech Synthesizer encompasses several potential avenues for further development and application. This includes the refinement of the emotion analysis model to enhance the accuracy and granularity of emotion identification within the input text. Additionally, expanding the system to support multiple languages beyond Malayalam would cater to a broader user base. Developing real-time emotion synthesis capabilities, allowing for on-the-fly conversion of neutral speech to emotive speech based on dynamic changes in the input text's emotional content, represents another area for future exploration. Furthermore, providing users with the ability to customize and fine-tune the emotional characteristics of the synthesized speech would enable personalized emotive speech generation. Integration with assistive technologies and devices, as well as the exploration of applications in diverse domains such as education, entertainment, and healthcare, would further expand the system's impact and practical implementation.

Chapter 7

References

(a) M. Divyapushpalakshmi, R. Ramalakshmi. An efficient sentimental analysis using hybrid deep learning and optimization technique for Twitter using parts of speech (POS) tagging. (2021).

(b) A.Naresh, P. Venkata Krishna. An efficient approach for sentiment analysis using machine learning algorithm. (2020).

(c) Kefei Cheng, Yanan Yue , And Zhiwen Song. Sentiment Classification Based On Part-of-speech And Self-attention Mechanism. (2020).

(d) Narisa Zhao , Huan Gao , Xin Wen , And Hui Li. Combination Of Convolutional Neural Network And Gated Recurrent Unit For Aspect-based Sentiment Analysis. (2021).

(e) Usha Devi Gandhi, Priyan Malarvizhi Kumar, Gokulnath Chandra Babu, Gayathri Karthick. Sentiment Analysis on Twitter Data by Using Convolutional Neural Network (CNN) and Long Short Term Memory (LSTM). (2021)

(f) Hossam Faris, Seyedali Mirjalili, Ibrahim Aljarah, Majdi Mafarja and Ali Asghar Heidari. Salp Swarm Algorithm: Theory, Literature Review, and Application in Extreme Learning Machines. (2020)

(g) Chen, J., Ye, L. and Ming, Z., 2021. Mass: Multi-task anthropomorphic speech synthesis framework. Computer Speech Language, 70, p.101243.

(h) Zhang, J., Wushouer, M., Tuerhong, G. and Wang, H., 2023. Semi-Supervised Learning for Robust Emotional Speech Synthesis with Limited Data. Applied Sciences, 13(9), p.5724.

(i) Elgaar, M., Park, J. and Lee, S.W., 2020, May. Multi-speaker and multi-domain emotional voice conversion using factorized hierarchical variational au-

toencoder. In ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 7769-7773). IEEE.

(j) Tits, N., 2019, September. A methodology for controlling the emotional expressiveness in synthetic speech-a deep learning approach. In 2019 8th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW) (pp. 1-5). IEEE.

(k) Tits, N., El Haddad, K. and Dutoit, T., 2020. Exploring transfer learning for low resource emotional tts. In Intelligent Systems and Applications: Proceedings of the 2019 Intelligent Systems Conference (IntelliSys) Volume 1 (pp. 52-60). Springer International Publishing.

(l) Abhash D, Sarmah P, Samudravijaya K, Prasanna SRM (2019) Development of Assamese text-to-speech system using deep neural network. In 2019 National Conference on Communications (NCC), pp. 1–5. IEEE.

(m) Absa AH, Deriche M, Elshafei-Ahmed M, Elhadj YM, Juang BH (2018) A hybrid unsupervised segmentation algorithm for Arabic speech using feature fusion and a genetic algorithm (July 2018). IEEE Access 6: 43157–43169.

(n) Li, Y.A., Zare, A. and Mesgarani, N., 2021. Starganv2-vc: A diverse, unsupervised, non-parallel framework for natural-sounding voice conversion. arXiv preprint arXiv:2107.10394.

Appendix A: Presentation

Emotive Malayalam Text-to-Speech: Vikara

By:
Akash Vijay
Aleena Mary Karatra
Alvin George Viji
Ashly Sabu

Guide:
Ms. Sherine Sebastian

CONTENTS:

PROBLEM DEFINITION

- There is a distinct lack of exploration in the field of Text to Speech Synthesis in the Malayalam Language .
- Moreover, current TTS systems lack emotion in the synthesized speech.
- However, most texts contain some emotion to be conveyed when converted to speech.
- The idea is to generate an Emotive Malayalam Text to Speech Synthesizer.
- The project aims to analyze the text for emotion, and generate the corresponding emotive speech for it.

PROJECT OBJECTIVE

1. To develop a Text-to-Speech Synthesizer for the Malayalam Language
2. To analyze the emotion of the input text using sentiment/emotion analysis
3. To generate emotion in the speech using emotion voice conversion after identifying the emotion in the text
4. To better the lives of the visually impaired, by providing expressive and emotive synthesized speech for Malayalam audiobooks, essays, and so on.

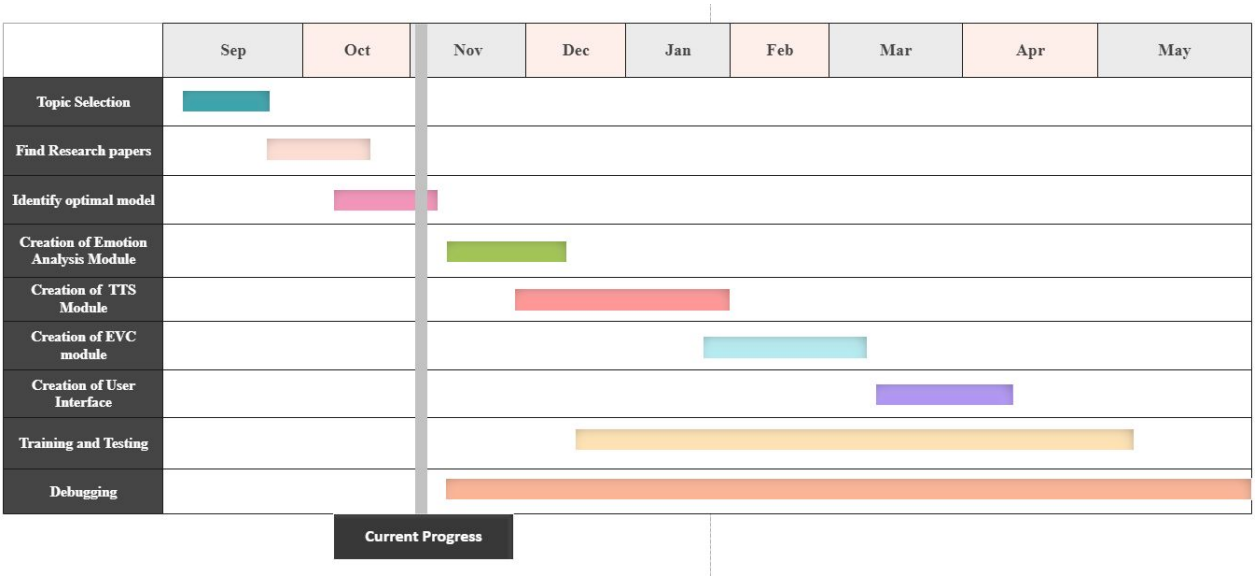
NOVELTY OF IDEA AND SCOPE OF IMPLEMENTATION

This technology aims to imbue synthesized speech with emotional nuances, allowing it to convey not just the literal meaning of the text but also the intended emotions behind it.

Key aspects are:

- Language Specificity
- Emotion Recognition

GANTT CHART



WORK DONE DURING 30% EVALUATION

Malayalam Text-to-Speech Engine

We created a malayalam TTS model, based on the Variational Inference with adversarial learning for end-to-end Text-to-Speech (VITS) model. The steps involved:

1. The text is sampled into the corresponding waveforms, ie, grapheme to phoneme conversion.
2. This sampling is done using a duration prediction module.
3. For the dataset, we were able to procure a dataset created by Facebook's multilingual venture.
4. The downloaded model with the pretrained checkpoint was called the "facebook/mms-tts-ml" for malayalam.

WORK PROGRESS (60% EVALUATION)

60% EVALUATION

For 60% evaluation we build a model using Logistic Regression for emotional analysis.

Steps :

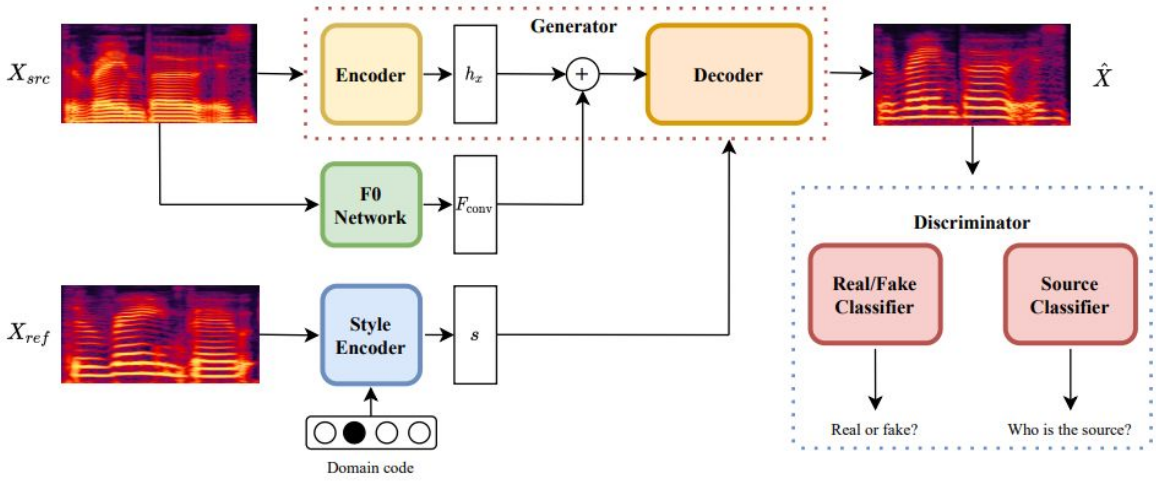
- Initially the emotion classes for each text in the dataset is identified and plotted
- Sentiments is analysed from each text and plotted
- Then the data is preprocessed by removing unwanted inputs
- Data is vectorized and split it into training and testing
- Dataset is trained using two different model (naive bayes and Linear Regression) and result is evaluated

- Use Google translation API for translating malayalam text to english
- The translated text is inputted to the model and the output is predicted

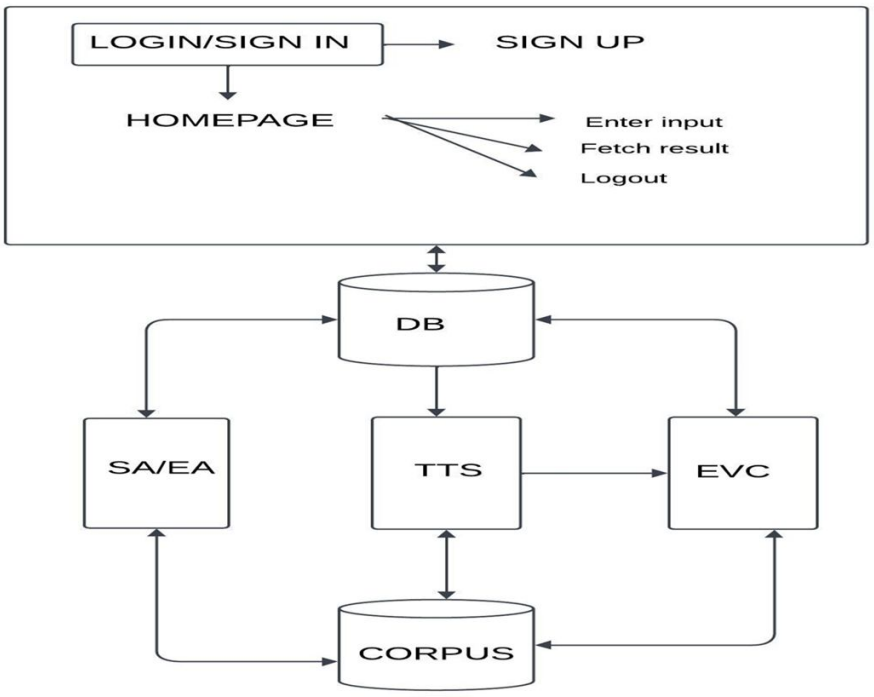
100% EVALUATION

Emotion Voice Conversion Module:

- Uses STARGANvc2 framework to convert the emotional characteristics of an audio sample
- Has a single discriminator and generator:
  - Generator: Converts input to required output, using style encoder(mel spectrogram) and F0 values
  - Discriminator: Binary Classifier that tries to classify whether a given sample is real or fake.
  - Mapping Network: To learn the acoustic characteristics of source and target speech and map them
- Dataset: Emotional Speech Dataset with 3500 audio samples and 10 different speakers, in both Mandarin and English.
- Has 5 emotions: Angry, Happy, Neutral, Sad and Surprise



Architecture Diagram of the Model:



INTERIM RESULTS

Malayalam Text to Speech conversion

```
text = "മലയാളം ഭാഷ ചരിത്രം."
inputs = tokenizer(text, return_tensors="pt")
```

```
from IPython.display import Audio
Audio(output, rate=model.config.sampling_rate)
```

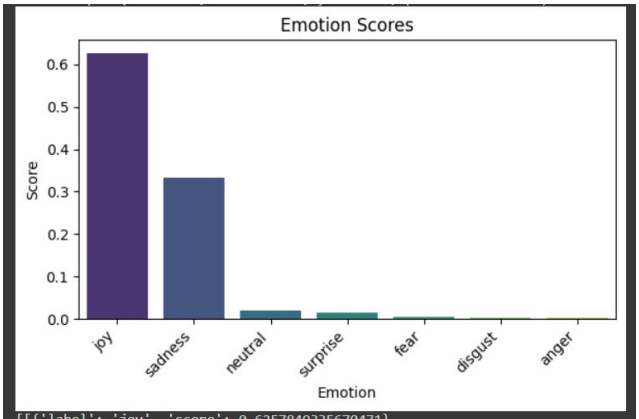
ittsmal.mp3

0:03 / 0:03

Emotional analysis

```
emotion_prediction("ഇന്ന് ഒരു മനോഹരമായ ദിവസമായിരുന്നു എന്നാൽ ഉച്ചയോടെ അത് മോശമായി തുടങ്ങി ")
```

```
Prediction: joy, Prediction Score: 0.6257849335670471
```




```
[[{'label': 'joy', 'score': 0.6257849335670471},
  {'label': 'sadness', 'score': 0.3325274884700775},
  {'label': 'neutral', 'score': 0.01873691938817501},
  {'label': 'surprise', 'score': 0.014042004942893982},
  {'label': 'fear', 'score': 0.004047735128551722},
  {'label': 'disgust', 'score': 0.0024961510207504034},
  {'label': 'anger', 'score': 0.00236484594643116}]]
```

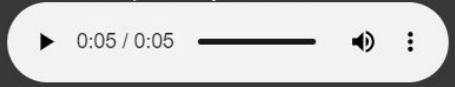
Uses a DistilRoBERTa-base Inference API for Emotional analysis

```
[ ] sample_text = "ഇന്ന് ഒരു മനോഹരമായ ദിവസമായിരുന്നു എന്നാൽ ഉച്ചയോടെ അത് മോശമായി തുടങ്ങി "
    emotion_prediction(sample_text)
```

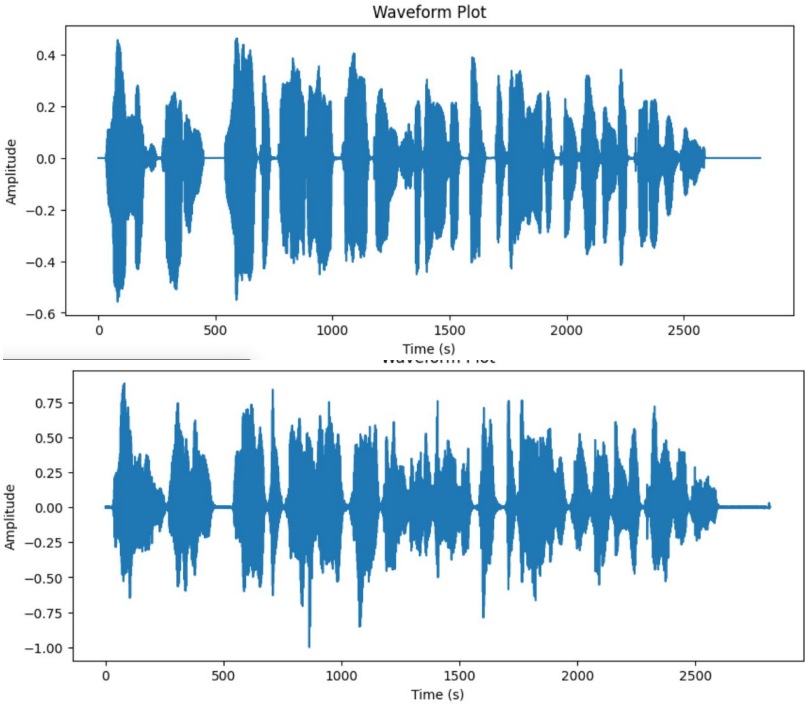
```
Prediction: joy, Prediction Score: 0.6257849335670471
```

```
output = TTS_audio(sample_text)

from IPython.display import Audio

Audio(output, rate=model.config.sampling_rate)
```

```
Some weights of the model checkpoint at facebook/mms-tts-mal were not used when initializing VitsModel: ['fl
- This IS expected if you are initializing VitsModel from the checkpoint of a model trained on another task
- This IS NOT expected if you are initializing VitsModel from the checkpoint of a model that you expect to b
Some weights of VitsModel were not initialized from the model checkpoint at facebook/mms-tts-mal and are new
You should probably TRAIN this model on a down-stream task to be able to use it for predictions and inferenc
```



Emotion Voice Conversion



Input

Output

Converted: 0/79.wav

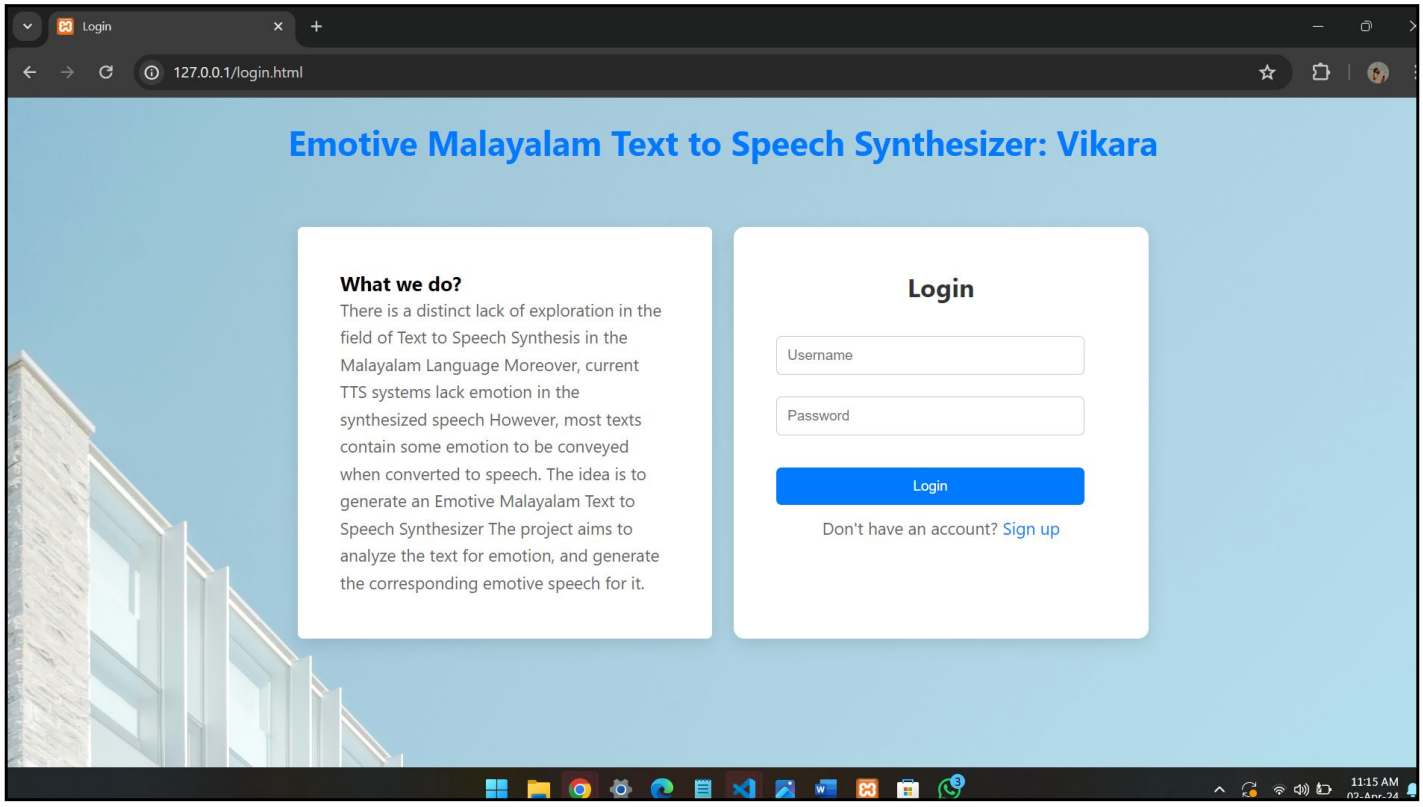
0:05 / 0:05

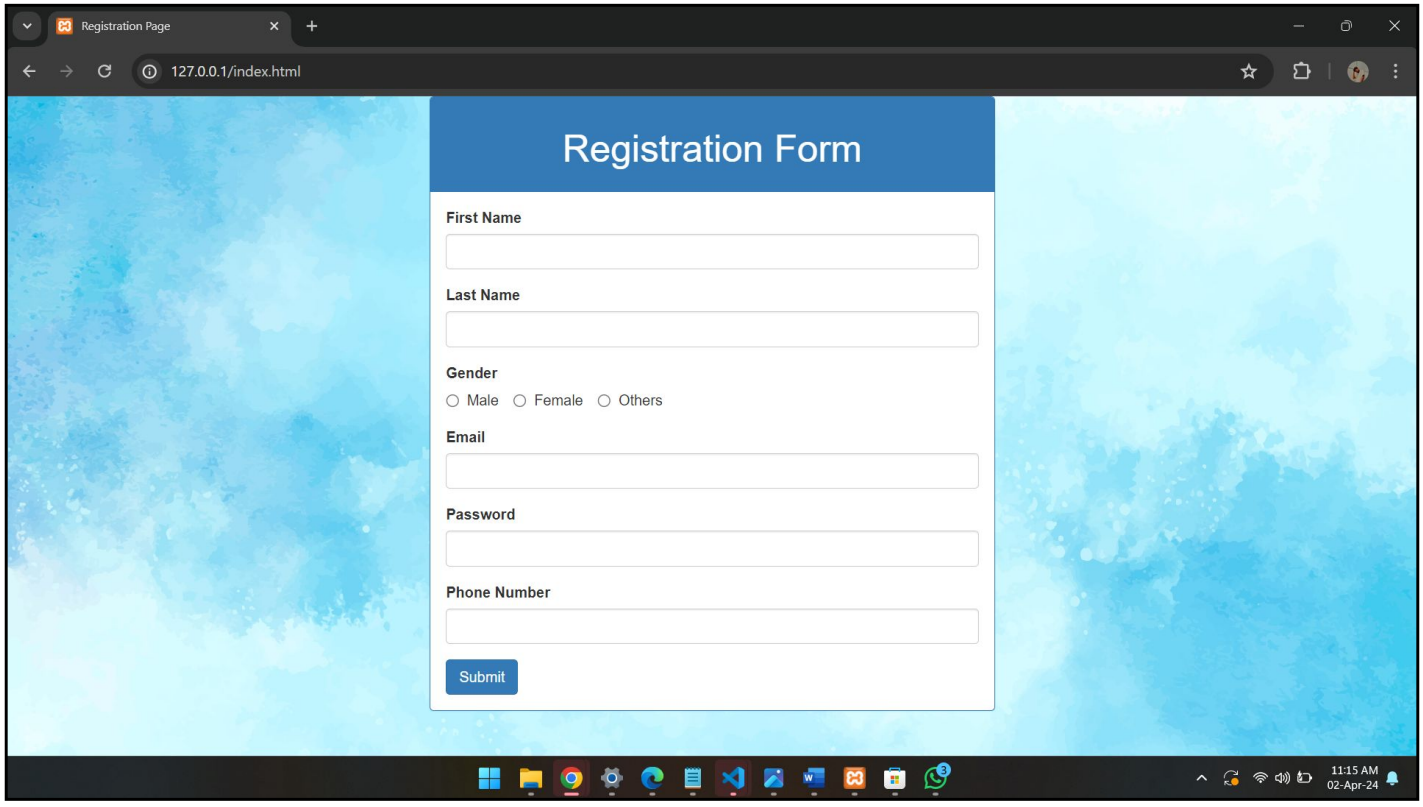
Reference (vocoder): 0/79.wav

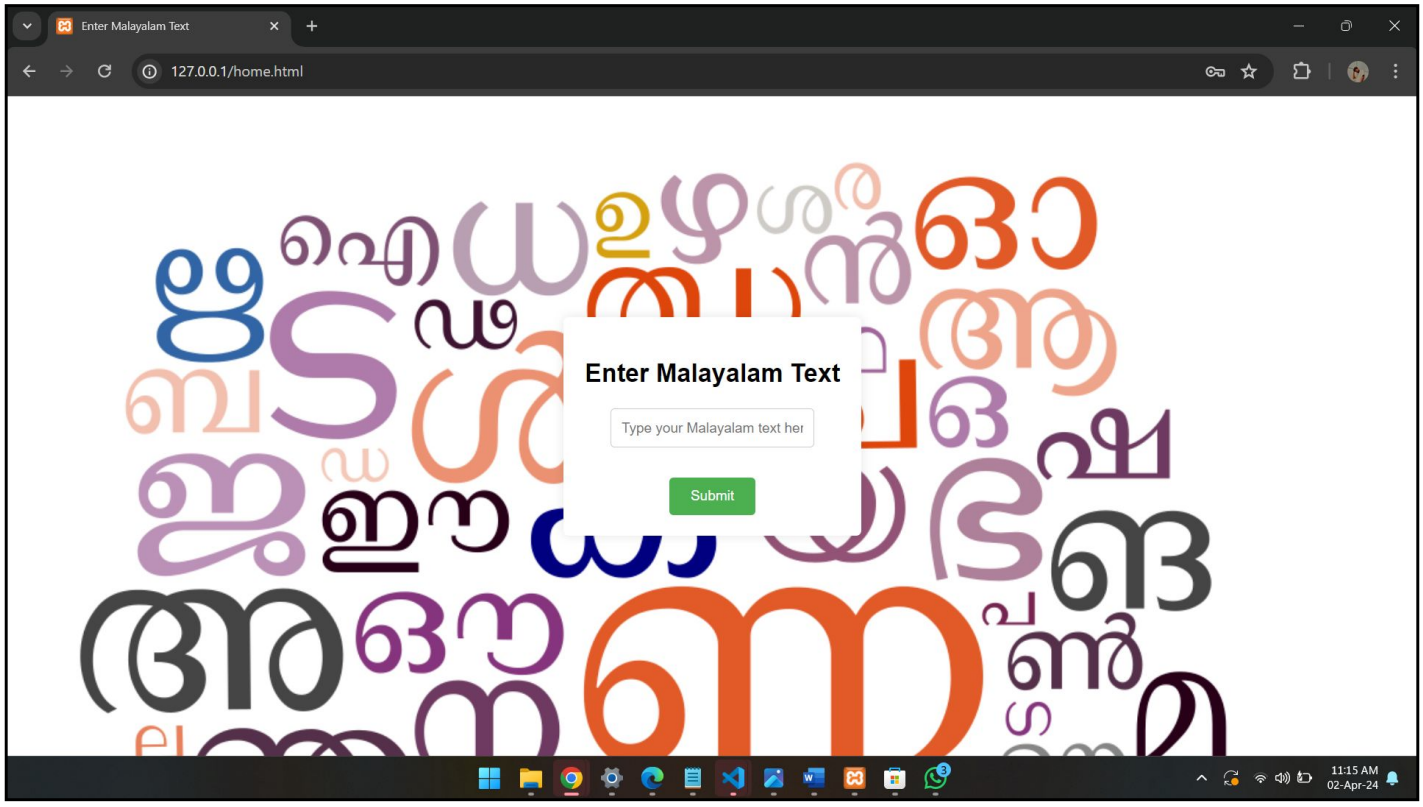
0:00 / 0:05

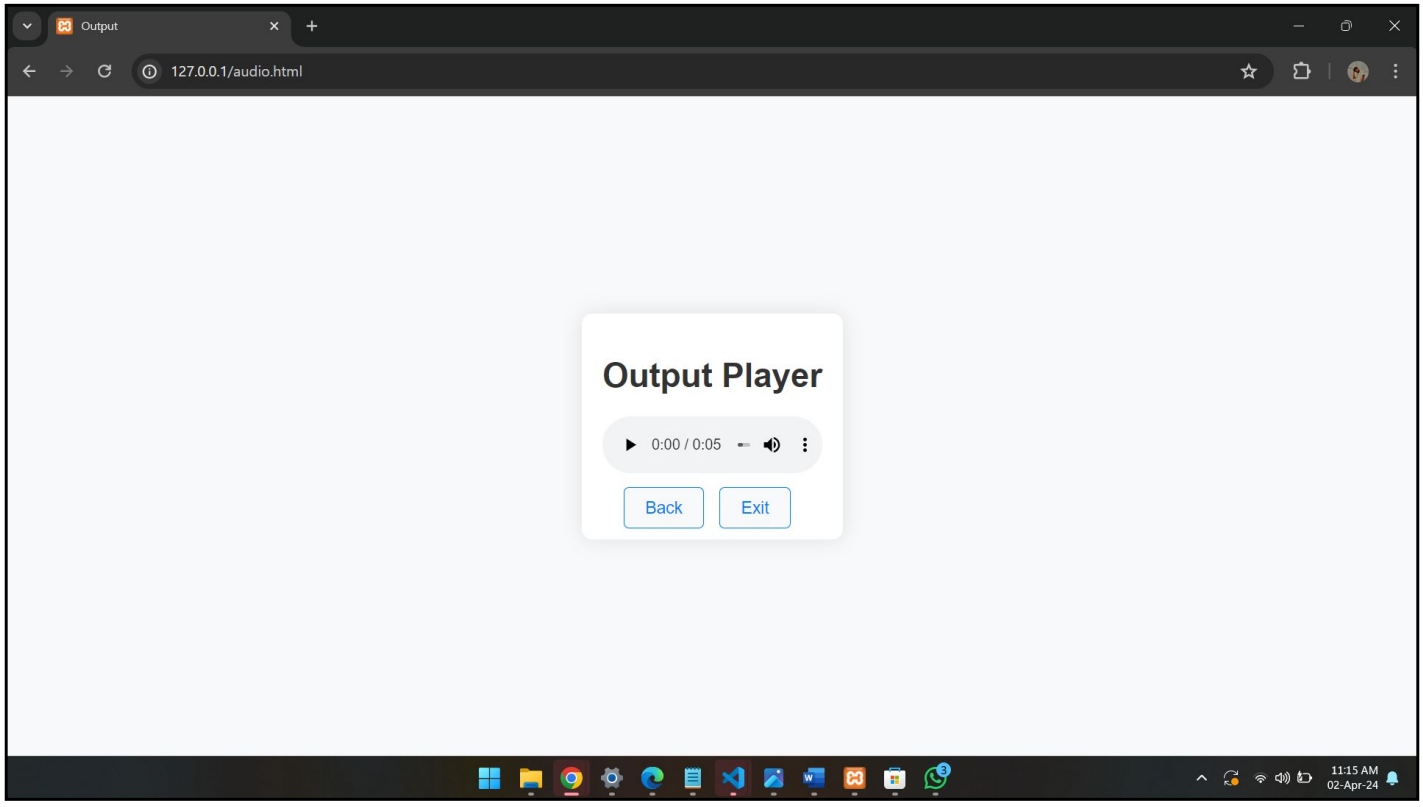
Sample.wav

Output.wav


Login
127.0.0.1/login.html
Emotive Malayalam Text to Speech Synthesizer: Vikara
What we do?
There is a distinct lack of exploration in the field of Text to Speech Synthesis in the Malayalam Language Moreover, current TTS systems lack emotion in the synthesized speech However, most texts contain some emotion to be conveyed when converted to speech. The idea is to generate an Emotive Malayalam Text to Speech Synthesizer The project aims to analyze the text for emotion, and generate the corresponding emotive speech for it.
Login
Username
Password
Login
Don't have an account? Sign up
11:15 AM
02-Apr-24


Registration Page
127.0.0.1/index.html
Registration Form
First Name
Last Name
Gender
Male
Female
Others
Email
Password
Phone Number
Submit
11:15 AM
02-Apr-24


Enter Malayalam Text
127.0.0.1/home.html
Enter Malayalam Text
Type your Malayalam text her
Submit
11:15 AM
02-Apr-24


Output
127.0.0.1/audio.html
Output Player
0:00 / 0:05
Back
Exit
11:15 AM
02-Apr-24

FUTURE SCOPE

- This project is currently, one of the only existing works for emotional voice synthesis for the Malayalam Language.
- However, due to lack of corpus availability, the project can be scaled to be used for a Malayalam dataset, like the currently unavailable, but soon to be published MAVES DB(Malayalam Audio-Visual Emotional Speech Data) for better emotional voice conversion.
- Similarly, the emotion identification module can be trained on a curated Malayalam corpus in the future.
- Speaker identity conversion can also be added to the project, to learn different accents and create a more immersive experience for the users, especially when it comes to converting dialogues of a novel.

TASK DISTRIBUTION

1. Akash: Emotion Analysis, Text-to-Speech Synthesis, Back-end
2. Aleena: (Lead) Emotion Voice Conversion, Front-end UI
3. Alvin: Text-to-Speech Synthesis, Emotion Analysis, Front end
4. Ashly: Emotion Voice Conversion, Emotion Analysis, Back-end

CONCLUSION

- Development of an emotive Malayalam Text to Speech (TTS) synthesis system involves:
  - Incorporating emotional expression into synthesized speech
  - Adding depth and nuance to the spoken output
- Applications span various domains:
  - Enhancing emotional impact in audio content
  - Creating engaging conversational agents
  - Providing a valuable tool for individuals with specific communication needs.

REFERENCES

1. M. Divyapushpalakshmi, R. Ramalakshmi. An efficient sentimental analysis using hybrid deep learning and optimization technique for Twitter using parts of speech (POS) tagging. (2021).
2. A.Naresh, P. Venkata Krishna. An efficient approach for sentiment analysis using machine learning algorithm. (2020).
3. Kefei Cheng, Yanan Yue , And Zhiwen Song. Sentiment Classification Based On Part-of-speech And Self-attention Mechanism. (2020).
4. Narisa Zhao , Huan Gao , Xin Wen , And Hui Li. Combination Of Convolutional Neural Network And Gated Recurrent Unit For Aspect-based Sentiment Analysis. (2021).
5. Usha Devi Gandhi, Priyan Malarvizhi Kumar, Gokulnath Chandra Babu, Gayathri Karthick. Sentiment Analysis on Twitter Data by Using Convolutional Neural Network (CNN) and Long Short Term Memory (LSTM). (2021)
6. Hossam Faris, Seyedali Mirjalili, Ibrahim Aljarah, Majdi Mafarja and Ali Asghar Heidari. Salp Swarm Algorithm: Theory, Literature Review, and Application in Extreme Learning Machines. (2020)
7. Chen, J., Ye, L. and Ming, Z., 2021. Mass: Multi-task anthropomorphic speech synthesis framework. Computer Speech & Language, 70, p.101243.
8. Zhang, J., Wushouer, M., Tuerhong, G. and Wang, H., 2023. Semi-Supervised Learning for Robust Emotional Speech Synthesis with Limited Data. Applied Sciences, 13(9), p.5724.

REFERENCES

9. Elgaar, M., Park, J. and Lee, S.W., 2020, May. Multi-speaker and multi-domain emotional voice conversion using factorized hierarchical variational autoencoder. In ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 7769-7773). IEEE.
10. Tits, N., 2019, September. A methodology for controlling the emotional expressiveness in synthetic speech-a deep learning approach. In 2019 8th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW) (pp. 1-5). IEEE.
11. Tits, N., El Haddad, K. and Dutoit, T., 2020. Exploring transfer learning for low resource emotional tts. In Intelligent Systems and Applications: Proceedings of the 2019 Intelligent Systems Conference (IntelliSys) Volume 1 (pp. 52-60). Springer International Publishing.
12. Abhash D, Sarmah P, Samudravijaya K, Prasanna SRM (2019) Development of Assamese text-to-speech system using deep neural network. In 2019 National Conference on Communications (NCC), pp. 1–5. IEEE.
13. 2. Absa AH, Deriche M, Elshafei-Ahmed M, Elhadj YM, Juang BH (2018) A hybrid unsupervised segmentation algorithm for Arabic speech using feature fusion and a genetic algorithm (July 2018). IEEE Access 6: 43157–43169.

REFERENCES

14. Afzal H Md, Memon S, Gregory MA (2010) A novel approach for MFCC features extraction, In 2010 4th International Conference on Signal Processing and Communication Systems, pp. 1–5. IEEE, 2010.

15. Ansal V (2020) ALO-optimized artificial neural network-controlled dynamic voltage restorer for compensation of voltage issues in distribution system. Soft Comput 24(2):1171–1184.

16. Archana B, Dev A, Kumari R, Agrawal SS (2016) Labelling of Hindi speech. IETE J Res 62ript to speech conversion for Hindi la(2):146–153.

Appendix B: Vision, Mission, Programme Outcomes and Course Outcomes

Vision, Mission, Programme Outcomes and Course Outcomes

Institute Vision
To evolve into a premier technological institution, moulding eminent professionals with creative minds, innovative ideas and sound practical skill, and to shape a future where technology works for the enrichment of mankind.

Institute Mission
To impart state-of-the-art knowledge to individuals in various technological disciplines and to inculcate in them a high degree of social consciousness and human values, thereby enabling them to face the challenges of life with courage and conviction.

Department Vision

To become a centre of excellence in Computer Science and Engineering, moulding professionals catering to the research and professional needs of national and international organizations.

Department Mission

To inspire and nurture students, with up-to-date knowledge in Computer Science and Engineering, ethics, team spirit, leadership abilities, innovation and creativity to come out with solutions meeting societal needs.

Programme Outcomes (PO)
Engineering Graduates will be able to:

1. Engineering Knowledge: Apply the knowledge of mathematics, science, engineering fundamentals, and an engineering specialization to the solution of complex

engineering problems.

2. Problem analysis: Identify, formulate, review research literature, and analyze complex engineering problems reaching substantiated conclusions using first principles of mathematics, natural sciences, and engineering sciences.

3. Design/development of solutions: Design solutions for complex engineering problems and design system components or processes that meet the specified needs with appropriate consideration for the public health and safety, and the cultural, societal, and environmental considerations.

4. Conduct investigations of complex problems: Use research-based knowledge including design of experiments, analysis and interpretation of data, and synthesis of the information to provide valid conclusions.

5. Modern Tool Usage: Create, select, and apply appropriate techniques, resources, and modern engineering and IT tools including prediction and modeling to complex engineering activities with an understanding of the limitations.

6. The engineer and society: Apply reasoning informed by the contextual knowledge to assess societal, health, safety, legal, and cultural issues and the consequent responsibilities relevant to the professional engineering practice.

7. Environment and sustainability: Understand the impact of the professional engineering solutions in societal and environmental contexts, and demonstrate the knowledge of, and need for sustainable development.

8. Ethics: Apply ethical principles and commit to professional ethics and responsibilities and norms of the engineering practice.

9. Individual and Team work: Function effectively as an individual, and as a member or leader in teams, and in multidisciplinary settings.

10. Communication: Communicate effectively with the engineering community and with society at large. Be able to comprehend and write effective reports documentation. Make effective presentations, and give and receive clear instructions.

11. Project management and finance: Demonstrate knowledge and understanding of engineering and management principles and apply these to one's own

work, as a member and leader in a team. Manage projects in multidisciplinary environments.

12. Life-long learning: Recognize the need for, and have the preparation and ability to engage in independent and lifelong learning in the broadest context of technological change.

Programme Specific Outcomes (PSO)

A graduate of the Computer Science and Engineering Program will demonstrate:

PSO1: Computer Science Specific Skills
The ability to identify, analyze and design solutions for complex engineering problems in multidisciplinary areas by understanding the core principles and concepts of computer science and thereby engage in national grand challenges.

PSO2: Programming and Software Development Skills
The ability to acquire programming efficiency by designing algorithms and applying standard practices in software project development to deliver quality software products meeting the demands of the industry.

PSO3: Professional Skills
The ability to apply the fundamentals of computer science in competitive research and to develop innovative products to meet the societal needs thereby evolving as an eminent researcher and entrepreneur.

Course Outcomes (CO)

Course Outcome 1: Model and solve real world problems by applying knowledge across domains (Cognitive knowledge level: Apply).

Course Outcome 2: Develop products, processes or technologies for sustainable and socially relevant applications (Cognitive knowledge level: Apply).

Course Outcome 3: Function effectively as an individual and as a leader in diverse teams and to comprehend and execute designated tasks (Cognitive knowledge level: Apply).

Course Outcome 4: Plan and execute tasks utilizing available resources within timelines, following ethical and professional norms (Cognitive knowledge level: Apply).

Course Outcome 5: Identify technology/research gaps and propose innovative/creative solutions (Cognitive knowledge level: Analyze).

Course Outcome 6: Organize and communicate technical and scientific findings effectively in written and oral forms (Cognitive knowledge level: Apply).

Appendix C: CO-PO-PSO Mapping

CO-PO AND CO-PSO MAPPING

	PO 1	PO 2	PO 3	PO 4	PO 5	PO 6	PO 7	PO 8	PO 9	PO 10	PO 11	PO 12	PSO 1	PSO 2	PSO 3
CO 1	2	2	2	1	2	2	2	1	1	1	1	2	3		
CO 2	2	2	2		1	3	3	1	1		1	1		2	
CO 3									3	2	2	1			3
CO 4					2			3	2	2	3	2			3
CO 5	2	3	3	1	2							1	3		
CO 6					2			2	2	3	1	1			3

3/2/1: high/medium/low

JUSTIFICATIONS FOR CO-PO MAPPING

MAPPING	LOW/ MEDIUM/HIGH	JUSTIFICATION
100003/ CS722U.1-PO1	M	Knowledge in the area of technology for project development using various tools results in better modeling.
100003/ CS722U.1-PO2	M	Knowledge acquired in the selected area of project development can be used to identify, formulate, review research literature, and analyze complex engineering problems reaching substantiated conclusions.

100003/ CS722U.1-PO3	M	Can use the acquired knowledge in designing solutions to complex problems.
100003/ CS722U.1-PO4	M	Can use the acquired knowledge in designing solutions to complex problems.
100003/ CS722U.1-PO5	H	Students are able to interpret, improve and redefine technical aspects for design of experiments, analysis and interpretation of data, and synthesis of the information to provide valid conclusions.
100003/ CS722U.1-PO6	M	Students are able to interpret, improve and redefine technical aspects by applying contextual knowledge to assess societal, health and consequential responsibilities relevant to professional engineering practices.
100003/ CS722U.1-PO7	M	Project development based on societal and environmental context solution identification is the need for sustainable development.
100003/ CS722U.1-PO8	L	Project development should be based on professional ethics and responsibilities.
100003/ CS722U.1-PO9	L	Project development using a systematic approach based on well defined principles will result in teamwork.
100003/ CS722U.1-PO10	M	Project brings technological changes in society.

100003/ CS722U.1-PO11	H	Acquiring knowledge for project development gathers skills in design, analysis, development and implementation of algorithms.
100003/ CS722U.1-PO12	H	Knowledge for project development contributes engineering skills in computing & information gatherings.
100003/ CS722U.2-PO1	H	Knowledge acquired for project development will also include systematic planning, developing, testing and implementation in computer science solutions in various domains.
100003/ CS722U.2-PO2	H	Project design and development using a systematic approach brings knowledge in mathematics and engineering fundamentals.
100003/ CS722U.2-PO3	H	Identifying, formulating and analyzing the project results in a systematic approach.
100003/ CS722U.2-PO5	H	Systematic approach is the tip for solving complex problems in various domains.
100003/ CS722U.2-PO6	H	Systematic approach in the technical and design aspects provide valid conclusions.
100003/ CS722U.2-PO7	H	Systematic approach in the technical and design aspects demonstrate the knowledge of sustainable development.

100003/ CS722U.2-PO8	M	Identification and justification of technical aspects of project development demonstrates the need for sustainable development.
100003/ CS722U.2-PO9	H	Apply professional ethics and responsibilities in engineering practice of development.
100003/ CS722U.2-PO11	H	Systematic approach also includes effective reporting and documentation which gives clear instructions.
100003/ CS722U.2-PO12	M	Project development using a systematic approach based on well defined principles will result in better teamwork.
100003/ CS722U.3-PO9	H	Project development as a team brings the ability to engage in independent and lifelong learning.
100003/ CS722U.3-PO10	H	Identification, formulation and justification in technical aspects will be based on acquiring skills in design and development of algorithms.
100003/ CS722U.3-PO11	H	Identification, formulation and justification in technical aspects provides the betterment of life in various domains.
100003/ CS722U.3-PO12	H	Students are able to interpret, improve and redefine technical aspects with mathematics, science and engineering fundamentals for the solutions of complex

		problems.
100003/ CS722U.4-PO5	H	Students are able to interpret, improve and redefine technical aspects with identification formulation and analysis of complex problems.
100003/ CS722U.4-PO8	H	Students are able to interpret, improve and redefine technical aspects to meet the specified needs with appropriate consideration for the public health and safety, and the cultural, societal, and environmental considerations.
100003/ CS722U.4-PO9	H	Students are able to interpret, improve and redefine technical aspects for design of experiments, analysis and interpretation of data, and synthesis of the information to provide valid conclusions.
100003/ CS722U.4-PO10	H	Create, select, and apply appropriate techniques, resources, and modern engineering and IT tools for better products.
100003/ CS722U.4-PO11	M	Students are able to interpret, improve and redefine technical aspects by applying contextual knowledge to assess societal, health and consequential responsibilities relevant to professional engineering practices.
100003/ CS722U.4-PO12	H	Students are able to interpret, improve and redefine technical aspects for demonstrating the knowledge of, and need for sustainable development.

100003/ CS722U.5-PO1	H	Students are able to interpret, improve and redefine technical aspects, apply ethical principles and commit to professional ethics and responsibilities and norms of the engineering practice.
100003/ CS722U.5-PO2	M	Students are able to interpret, improve and redefine technical aspects, communicate effectively on complex engineering activities with the engineering community and with society at large, such as, being able to comprehend and write effective reports and design documentation, make effective presentations, and give and receive clear instructions.
100003/ CS722U.5-PO3	H	Students are able to interpret, improve and redefine technical aspects to demonstrate knowledge and understanding of the engineering and management principle in multidisciplinary environments.
100003/ CS722U.5-PO4	H	Students are able to interpret, improve and redefine technical aspects, recognize the need for, and have the preparation and ability to engage in independent and life-long learning in the broadest context of technological change.
100003/ CS722U.5-PO5	M	Students are able to interpret, improve and redefine technical aspects in acquiring skills to design, analyze and develop algorithms and implement those using high-level programming languages.
100003/ CS722U.5-PO12	M	Students are able to interpret, improve and redefine

		technical aspects and contribute their engineering skills in computing and information engineering domains like network design and administration, database design and knowledge engineering.
100003/ CS722U.6-PO5	M	Students are able to interpret, improve and redefine technical aspects and develop strong skills in systematic planning, developing, testing, implementing and providing IT solutions for different domains which helps in the betterment of life.
100003/ CS722U.6-PO8	H	Students will be able to associate with a team as an effective team player for the development of technical projects by applying the knowledge of mathematics, science, engineering fundamentals, and an engineering specialization to the solution of complex engineering problems.
100003/ CS722U.6-PO9	H	Students will be able to associate with a team as an effective team player to Identify, formulate, review research literature, and analyze complex engineering problems
100003/ CS722U.6-PO10	M	Students will be able to associate with a team as an effective team player for designing solutions to complex engineering problems and design system components.
100003/ CS722U.6-PO11	M	Students will be able to associate with a team as an effective team player, use research-based knowledge and research methods including design of experiments, analysis and interpretation of data.

100003/ CS722U.6-PO12	H	Students will be able to associate with a team as an effective team player, applying ethical principles and commit to professional ethics and responsibilities and norms of the engineering practice.
100003/ CS722U.1-PSO1	H	Students are able to develop Computer Science Specific Skills by modeling and solving problems.
100003/ CS722U.2-PSO2	M	Developing products, processes or technologies for sustainable and socially relevant applications can promote Programming and Software Development Skills.
100003/ CS722U.3-PSO3	H	Working in a team can result in the effective development of Professional Skills.
100003/ CS722U.4-PSO3	H	Planning and scheduling can result in the effective development of Professional Skills.
100003/ CS722U.5-PSO1	H	Students are able to develop Computer Science Specific Skills by creating innovative solutions to problems.
100003/ CS722U.6-PSO3	H	Organizing and communicating technical and scientific findings can help in the effective development of Professional Skills.