DEEPFAKE TEXT TO SPEECH (TTS)

Final Year Project – Final Report

Session 2020-2024

**A 4th Year Student**

A project submitted in partial fulfilment of the

COMSATS University Degree

of Computer Science

BSc. (Hons.)BS in Computer Science Engineering (CUI)



Department of Computer Science

COMSATS University Islamabad, Lahore Campus

# Project Detail

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Type (Nature of project) | | | [ **X** ] **D**evelopment [ ] **R**esearch [ ] **R**&**D** | | |
| Area of specialization | | | Deepfakes audio generation | | |
| **Project Group Members** | | | | | |
| Sr.# | Reg. # | Student Name | | Email ID | \*Signature |
| (i) | FA20-BCS-007 | Muhammad Talha | | Fa20-bcs-007@cuilahore.edu.pk |  |
| (ii) | SP20-BCS-114 | Hassan Mahmood | | Sp20-bcs-114@cuilahore.edu.pk |  |

\*The candidates confirm that the work submitted is their own and appropriate credit has been given where reference has been made to work of others

Supervisor Name: **Dr. Wajahat Mehmood Qazi** Supervisor Signature: \_\_\_\_\_\_\_\_\_\_\_\_\_\_

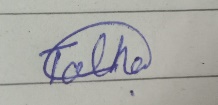
Remarks (if any): \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

# Plagiarism Free Certificate

This is to certify that, I am **Muhammad Talha** S/D/o **Falak Sher,** group leader of FYP under registration no CIIT/ **FA20-BCS-007** /LHR at Computer Science Department, COMSATS Institute of Information Technology, Lahore. I declare that my FYP proposal is checked by my supervisor and the similarity index is \_\_\_\_\_\_\_\_% that is less than 20%, an acceptable limit by HEC. Report is attached herewith as Appendix A.

Date: \_\_\_\_\_\_\_\_\_\_\_\_ Name of Group Leader: **Muhammad Talha** Signature: 

Name of Supervisor: **Dr Wajahat Mehmood Qazi** Co-Supervisor (if any):\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Designation: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Designation: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Signature: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Signature: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

HoD: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Signature: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Abstract

This project aims to develop an application that can generate voice clones based on provided text using a trained model, we named this app **Deepcloning.ai**. Many Text-to-Speech models exist, they are limited to specific voices. Our focus is on training a model with an English dataset to replicate a particular person's voice. Once trained, the model can produce Deepfakes voices for the provided text, offering advantages in education, cost-effective advertising, and the media industry. However, it's essential to acknowledge the potential for misuse, especially in serious crimes and impersonation, necessitating subsequent efforts for detection and prevention.This project's primary objective is Deepfakes text to speech creation, with ethical and security considerations in mind.

Table of Contents

[Project Detail 2](#_Toc164339579)

[Plagiarism Free Certificate 2](#_Toc164339580)

[1 Introduction 8](#_Toc164339581)

[1.1 Introduction 8](#_Toc164339582)

[1.2 Goals and Objectives 9](#_Toc164339583)

[1.3 Problem Statement 10](#_Toc164339584)

[1.4 Assumptions and Constraints 10](#_Toc164339585)

[1.5 Project Scope 11](#_Toc164339586)

[2 Requirement Analysis 14](#_Toc164339587)

[2.1 Literature Review 14](#_Toc164339588)

[2.1.1 Literature Review: Audio Deepfakes: A survey by Zahra Khanjani, Gabrielle Watson and Vandana P. Janeja\* [5] 14](#_Toc164339589)

[2.1.2 Existing System Review: Retrieval-Based Voice Conversion (RVC) vocal model. [6] 16](#_Toc164339590)

[2.1.3 Existing System Review: XTTS (experimental TTS) and Tortoise TTS from Coqui. 18](#_Toc164339591)

[2.2 Stakeholders list: 18](#_Toc164339592)

[2.3 Product Functions: 19](#_Toc164339593)

[2.3.1 User characteristics: 20](#_Toc164339594)

[2.4 Requirements Elicitation: 20](#_Toc164339595)

[2.4.1 Functional requirements: 21](#_Toc164339596)

[2.4.2 Non-Functional requirements: 23](#_Toc164339597)

[2.5 Use Case Description: 26](#_Toc164339598)

[2.6 Software Development Life Cycle: 33](#_Toc164339599)

[3 System Design: 36](#_Toc164339600)

[3.1 Work Breakdown Structure: (WBS) 36](#_Toc164339601)

[3.2 Activity Diagram: 37](#_Toc164339602)

[3.2.1 Login: 37](#_Toc164339603)

[3.2.2 Text to Speech: 38](#_Toc164339604)

[3.2.3 Voice Cloning: 39](#_Toc164339605)

[3.2 Sequence Diagram: 40](#_Toc164339606)

[3.2.1 Login: 40](#_Toc164339607)

[3.2.2 Text to Speech 41](#_Toc164339608)

[3.2.3 Voice Cloning 42](#_Toc164339609)

[3.3 Software Architecture: 43](#_Toc164339610)

[3.4 Collaboration Diagram 44](#_Toc164339611)

[4 System Testing 46](#_Toc164339612)

[4.1 Test Cases 46](#_Toc164339613)

[4.2 Unit Testing 49](#_Toc164339614)

[4.3 Acceptance Testing 49](#_Toc164339615)

[5 Application Frontend 51](#_Toc164339616)

[5.1 Web Interface 51](#_Toc164339617)

[5.1.1 Login Screen 51](#_Toc164339618)

[5.1.2 Sign Up Screen 52](#_Toc164339619)

[5.1.3 Home Screen 53](#_Toc164339620)

[5.1.4 Text to Speech 54](#_Toc164339621)

[5.1.5 Voice Cloning 55](#_Toc164339622)

[5.1.6 Contact Us 56](#_Toc164339623)

[6 Conclusion 58](#_Toc164339624)

[6.1 Problem faced 58](#_Toc164339625)

[References 61](#_Toc164339626)

Chapter 1: Introduction

# Introduction

## Introduction

This project is an **industrial project** and industry name is **Expert System Solution.** Deepfakes is a combination of words Deep which is Deep learning and Fake. Deep Learning is basically Artificial Intelligence based Learning Architecture which is trained on Dataset and give you a desire results. Creating something fake using Deep Learning is basically a Deepfakes. While the act of creating fake content is not new, deep fakes leverage powerful techniques from machine learning and artificial intelligence to manipulate or generate visual and audio content that can more easily deceive. Deep fakes started impacting society in various ways like manipulation of facial appearance, Voice Cloning of famous Personalities, Fake Text Generation. [1]

Deep learning has made significant progress in recognizing actions in videos and Voice in real-time, and it no longer requires expensive and powerful servers as it used to. Thanks to improvements in hardware and software, we can now use these deep learning models on everyday devices like mobile phone. This development has opened exciting possibilities for creation of Deepfakes Content, which our project aims to leverage fully. The most common deep learning-based machine learning techniques for producing Deepfakes include training generative neural networks designs like auto encoders or generative adversarial networks (GANs).[2]

Deepfakes are manufactured works of art in which the resemblance of a different person is used to replace a real person in a photograph, video and Voice. Now days, it is use in News Media as India channel Odisha TV launches the Deepfakes audio and animation-based News Anchor. [3]Ukraine President Deepfakes video goes viral which is in support of Russia cause a chaos in Public.[4] Now days, Deepfakes is very much in use, mostly for Negativity in the society such as Harassment, False Information, the use of deep fakes since they began in 2017 has been shocking. A huge increase in articles related to Deepfakes have been noticed from 2018 and in 2020, it was expected that around 730 articles will be publish related to Deepfakes but in reality at end of year number was way more than that around 1323 [5], which shows the interest of people in exploring this field.

This Project aims the creation of Deepfakes text to speech on a single person dataset, for this purpose considered datasets given below. In future, this project will provide us help in Deepfakes detection and prevention as well as expanding the scope for multiple person’s Deepfakes and personal voice Deepfakes**.**

## Goals and Objectives

The aim of this project is to develop a Deepfakes text-to-speech system for voice cloning, initially focusing on cloning the voice of one specific person and subsequently expanding to a broader range of individuals.

1. Achieving satisfactory accuracy in our model training on custom datasets.
2. Ensuring that the voice cloning can deceive the audience
3. State of the art custom dataset of a specific person which can to develop a realistic Deepfakes.

**Objectives:**

1. Create a Deepfakes voice generation model capable of producing convincing audio from text input.

2. Address the ethical concerns associated with voice cloning, particularly in legal and law enforcement contexts.

3. Construct a comprehensive dataset for training Deepfakes voice generation models.

4. Lay the groundwork for future research in Deepfakes voice detection.

6. Promote responsible and ethical use of Deepfakes voice technology.

## Problem Statement

In today's world, it's getting harder to tell real audio from fake. Some people use computers to copy famous voices, like actors or politicians, and make them say things they never actually said. This is a big problem because it leads to lies, scams, and confusion. We're facing a growing issue with Deepfakes text-to-speech voice cloning. It means that people can copy someone's voice, especially famous folks, using fancy computer tricks. This has some serious consequences, like the risk of identity theft, spreading false information, and using audio in tricky ways. This is a significant problem because it can result in fake news stories that make it seem like famous people are saying things they never said. Crooks can also use this to trick people over the phone and steal their money. It's like someone pretending to be you by using your voice without your permission. It can also damage someone's reputation by making them look like they support things they don't.

Our project is all about addressing these challenges. We aim to enhance the quality of computer-generated voices, making them sound as realistic as possible. We're working on advancing the technology to create synthetic voices that are indistinguishable from real ones. We're committed to using this technology responsibly and ensuring that users understand its proper use. Furthermore, we'll closely monitor technological developments to remain effective and responsible in a world where fake audio is becoming a significant concern.

## Assumptions and Constraints

To effectively address the problem statement, several assumptions and constraints need to be considered:

**Voice Quality**: We assume that the quality of generated voices will be close enough to pass as authentic and that it will be challenging to differentiate between genuine and synthetic voices.

**Textual Input**: The system assumes that textual input will be available for generating synthetic voices. This text should be clear and coherent to produce realistic speech.

**Training Data**: The success of voice generation relies on access to a substantial amount of audio data from the targeted famous individuals. The quality and quantity of this data are essential for training the voice models.

**Privacy and Consent**: The project assumes that the usage of the generated voices will adhere to legal and ethical standards. Obtaining consent for voice cloning, especially for famous personalities, will be an important ethical consideration.

Constraints:

**Voice Duration**: Our model is constrained to generate audio that is up to 2 mins in length. Generating longer voice clips will require significant modifications and may not be feasible within the current project scope.

**Limited Non-Speech Sounds:** We have a limited library of non-speech sounds, including [laughter], [laughs], [sighs], [music], [gasps], and [clears throat]. Expanding this library with additional sounds may pose challenges due to resource and time constraints.

**Computational Resources:** The project is constrained by the available computational resources, which may affect the speed and efficiency of voice generation.

**Technological Constraints**: The project may be subject to technological limitations in terms of hardware, software, and computational resources. Availability and access to high-performance hardware may be a constraint.

## Project Scope

The project scope involves both the development of the voice cloning system. It encompasses the following aspects:

**Voice Generation Model**: The project includes the development and training of a voice generation model that can produce synthetic voices based on textual input. The scope covers acquiring and processing training data, model development, and optimization.

**Voice Cloning:** The system aims to clone the voices of famous individuals, but it will not address ethical or legal concerns related to voice cloning. It assumes that voice cloning will be done with consent and within legal boundaries.

**User Interface:** The project will have a user-friendly interface for users to interact with the system, both for voice generation and voice cloning detection

**Testing and Evaluation**: The project scope involves rigorous testing and evaluation of the voice generation components to ensure their accuracy and reliability.

**Documentation and Reporting**: The team will produce comprehensive documentation and reports, including guidelines on ethical usage, for end-users and relevant stakeholders.

**Ongoing Maintenance:** The project scope may also encompass ongoing maintenance and updates to address emerging challenges in voice cloning and Deepfakes technologies.

Chapter 2: Requirement Analysis

# Requirement Analysis

## Literature Review

### Literature Review: Audio Deepfakes: A survey by Zahra Khanjani, Gabrielle Watson and Vandana P. Janeja\* [5]

Audio Deepfakes: A survey, this literature gives a brief concept of Deepfakes and its working. First part of the literature describes what Deepfakes are and how Artificial Intelligence help us in developing Deepfakes. This paper explains the concept of Deepfakes as it’s an AI-generated or manipulated content, including audio, video, images, or text, designed to appear as authentic artefacts. Distinguished from manual editing, Deepfakes closely resembles real-life elements. Their impact is evident in politics, social media, and entertainment. Further this paper explains four types of Ai Technologies for Deepfakes generation: encoder–decoder networks (ED), convolutional neural networks (CNN), generative adversarial networks (GAN), and recurrent neural networks (RNN). This paper categorizes Deepfakes into four types: audio, text, video, and image Deepfakes. This paper focuses on providing a survey related to each category and briefly discussing the technology trends and frameworks. But notably, the paper emphasizes that audio Deepfakes have been overlooked in previous surveys, making this paper the first to specifically focus on generating and detecting audio Deepfakes. This paper delves into important frameworks, offering readers detailed insights and guidance on tools for audio Deepfakes creation and detection.

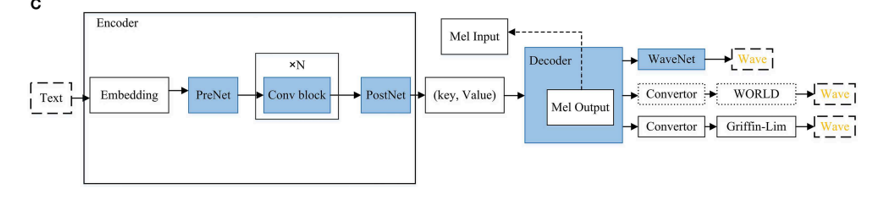
**Audio Deepfakes:**

Major working of this paper is on audio Deepfakes. Audio Deepfakes are AI-generated designed to sound authentic. Further I explain the types of audio Deepfakes include text-to-speech and voice conversion, including impersonation. Further Paper distinguishes between non-AI-generated audio fakes, such as replay attacks, and AI-generated or audio Deepfakes. It also explains how Deep Learning models such as Convolution Neural Network Help us in Detection of Replay Attacks. This paper further focuses on development of audio Deepfakes application and its challenges as developing effective speech synthesis systems demands significant processing power, storage, and high-quality speech corpora. Challenges include the expense of corpus creation, difficulties in modification, and issues with sparsely spoken languages. Rapid Advancement in this field making it easier for AI generated Deepfakes will be having a huge impact which would be difficult for detection.

* **Text-To-Speech:**

This paper briefly explains how audio Deepfakes using TTS converts text input into human-like speech. TTS offers advantages such as creating speech from scratch, supporting different accents and voices, and allowing configuration of aspects like speaking rate and pitch. TTS can be performed using various generative networks (GANs), and the text introduces frameworks like Char2Wav, WaveNet, WaveGlow and Tacotron.

**Tacotorn TTS (Coqui.AI):**

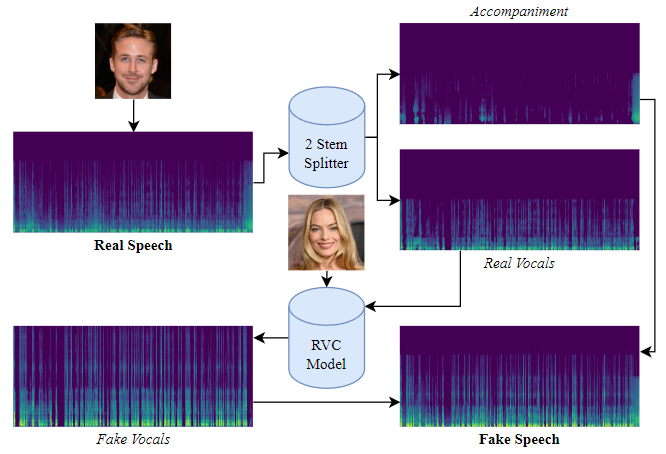


* **Voice Conversion and impersonate:**

Another subcategory of audio Deepfakes explain in paper which is known as voice conversion explained briefly, which modifies speech from a source speaker to sound like a target speaker. It outlines the positive applications of voice conversion, such as aiding those with speech disorders. The Literature introduces various voice conversion methods, including traditional models like Gaussian mixture models and modern approaches involving deep learning.

### Existing System Review: Retrieval-Based Voice Conversion (RVC) vocal model. [6]

The great advancement witnessed in voice cloning technology using Artificial Intelligence (AI) with the advent of **RVC** (Retrieval-Based Voice Conversion) Vocal Models. These models signify a notable advancement in voice cloning and text-to-speech (TTS). The RVC v2 model capable of developing high-quality, realistic AI voices.. Figure given below, shows the working of this model for Deepfakes voice cloning.



The effective voice cloning of your target person voice using RVC Vocal Models lies in comprehensive dataset use for model training. The dataset of audio should include wide range of voices and speech patterns. Voice should be clear and having no background noise. More importantly, small gaps of silence in speech and should be in proper English. The training process, often involving 60 to 150 epochs, is conducted on powerful GPUs to ensure efficiency and speed.

A proper guide provided for dataset preparation for training model by RVC and they also share their repository on GitHub as it’s an open source model. [7]

### Existing System Review: XTTS (experimental TTS) and Tortoise TTS from Coqui.

Tortoise is a text-to-speech (TTS) system known for its expressiveness and ability to clone voices. Here's a breakdown of its key features. Having features like Multi-Voice Capable, Realistic Prosody & Intonation and Voice Cloning. Architecture: It combines three different models: An autoregressive acoustic model that converts text into codes representing sounds. A Univnet vocoder that converts the mel spectrogram into the final audio waveform. A diffusion model that transforms these codes into a mel spectrogram, a visual representation of speech.

Tortoise excels at **voice cloning**. You can provide audio samples of a specific speaker. **Emotional Nuance:** Through "prompt engineering," you can influence the emotional tone of the speech by adding prompts within the text.

Tortoise TTS is a **Slower** model and need resources like GPU for usage. Where as **XTTS** is a super cool Text-to-Speech model that lets you clone voices in different languages built on the Tortoise TTS. Behind the scene, it uses Tortoise TTS with multi-languages support and have changes in working of Architecture.

That’s why it has lot of Advantages as there is no need for an excessive amount of training data. The time required for generating results are much lesser compare to Tortoise TTS. It’s the same model that powers [Coqui Studio](https://coqui.ai/), and [Coqui API](https://docs.coqui.ai/docs). As, both models are developed by the same company Coqui and it’s like a new version of Tortoise TTS.

## 2.2 Stakeholders list:

* Project Team
* Project Supervisor
* Available voices of persons for Cloning
* Project Owner

## 2.3 Product Functions:

Deepfakes text to speech consist of these following component:

**Web Application:** The web application will serve as a web-based interface for the end users of Deepfakes text to speech, providing them with a cluster of functionalities designed for the system such as Text to speech and voice cloning.

Deepfakes text to speech provides the following functionality to the user:

* **Login**: Log into the system using registered credentials or signup into application using email, name and password.
* **Text to Speech:** Our Product provide Text to Speech functionality. Which provide user to write text in text box in application then select the language of text provided. Afterward, user can select available voices related to language selected. Then can generate output which will be a cloned voice of a target person.
* **Voice Cloning:** This Product provide voice cloning functionality. Which provide user to record audio using microphone or import audio from device in application then select the language of audio provided. Afterward, user can select available voices related to language selected. Then can generate output which will be a cloned voice of a target person.

### 2.3.1 User characteristics:

The types of users identified that can use this system are:

* User of all type except the misusers

User Characteristics

*Table 01 (User Characteristics)*

|  |  |  |  |
| --- | --- | --- | --- |
| **User** | **Level of Computer Knowledge** | **Level of Business Knowledge** | **Frequency of Use** |
| User | Basic to advance knowledge | Can easily follow the system’s instructions | Constantly Using over an extended period of time |

In the above Table 01, a detailed description of the user of the system is described.

## 2.4 Requirements Elicitation:

As it’s an industrial project, so we get the requirements but through our observation and brainstorming with expert of AI machine learning field, we finalize the requirements. For our technical requirements, we read a research paper and explore relative Applications like Revoicer.ai [8], Speechify.ai [9]. Further we study its use cases and its purpose of integration with other applications.

### 2.4.1 Functional requirements:

Functional requirements are the Functionality expected from the system. Functional Requirement are written in more easy and natural language so that any layperson can read them and get a good understanding of the functionality of the system.

.

**FR-01: Login**

*Table 02 (FR01 - Log in)*

|  |  |
| --- | --- |
| **Req. No.** | **Functional Requirements** |
| **FR01-01** | The system shall grant the user access after verification of login credentials |
| **FR01-02** | The system shall allow the user to change his/her username and password |
| **FR01-03** | The system shall maintain a database of all the user accounts |

In the above Table 02, the functional requirements of login are mentioned.

**FR-02: Text-to-Speech (TTS)**

*Table 03 (FR02 – Text-to-Speech)*

|  |  |
| --- | --- |
| **Req. No.** | **Functional Requirements** |
| **FR02-01** | The system shall provide user a Text Box for writing text for converting to speech. |
| **FR02-02** | The system shall provide language English for Text. |
| **FR02-03** | The system shall provide the pre trained target person voices according to language selected. |
| **FR02-04** | The system shall show the generated voice output. |
| **FR02-05** | The system shall show an appropriate message for error or if text is not acceptable. |

In the above Table 03, the functional requirements of Text-to-speech are mentioned.

**FR-03: Voice Cloning Speech-to-Speech**

*Table 04 (FR03 – voice cloning)*

|  |  |
| --- | --- |
| **Req. No.** | **Functional Requirements** |
| **FR03-01** | The system should first ask user to login to access this feature. |
| **FR03-02** | The system shall provide user a microphone access option or with an audio file import option for having an audio file for voice cloning. |
| **FR03-03** | The system shall provide the pre trained target person voices according to language selected. For example, if user select English, then available voices of English will be shown. |
| **FR03-04** | The system shall show the generated voice output. |
| **FR03-05** | The system shall show an appropriate message for error or if text is not acceptable. |

In the above Table 04, the functional requirements of voice cloning are mentioned.

### 2.4.2 Non-Functional requirements:

Non-Functional requirements set the standards for the system. They define what is expected from the system apart from its basic functionality. They are required as a measure of quality check during testing and cover areas such as Performance, Usability, Reusability, Security, Reliability, Compatibility, Understanding, etc.

#### NFR-01: Performance

*Table 05 (NFR01 - Performance)*

|  |  |
| --- | --- |
| **Req. No.** | **Non-Functional Requirement** |
| **NFR01-1** | The system should generate the output in 2-3 minutes. |
| **NFR01-2** | The start-up time of the system must not be more than 45 seconds. |

In the above Table 05, non-functional requirements of the software's performance are mentioned.

#### NFR-02: Usability

*Table 06 (NFR02 - Usability)*

|  |  |
| --- | --- |
| **Req. No.** | **Non-Functional Requirement** |
| **NFR02-1** | The User must be able to get familiar with the interface no more than 5 minutes. |
| **NFR02-2** | The System must have minimal design with user-friendly interface |

In the above Table 06, non-functional requirements of the software’s usability are mentioned.

#### NFR-03: Reusability

*Table 07 (NFR 03 - Reusability)*

|  |  |
| --- | --- |
| **Req. No.** | **Non-Functional Requirement** |
| **NFR03-1** | The system must allow the reuse of functions in the system in a different environment |

In the above Table 07, non- functional requirements of the software’s reusability are mentioned.

#### NFR-04: Defect Maintenance

*Table 09 (NFR 04 – Defect Maintenance)*

|  |  |
| --- | --- |
| **Req. No.** | **Non-Functional Requirement** |
| **NFR04-1** | After implementation, the system must produce a minimum about of errors |
| **NFR04-2** | Debugging should not take more than 1 to 2 days |

#### In the above Table09, non-functional requirements of the software’s defect maintenance are mentioned.

#### NFR-05: Completeness

*Table 10 (NFR05 - Completeness)*

|  |  |
| --- | --- |
| **Req. No.** | **Non-Functional Requirement** |
| **NFR05-1** | The System must always be consistent in generating nearly accurate output (cloned voice). |

In the above Table 10, non-functional requirements of the software’s completeness are mentioned

#### NFR-06: Security

*Table 11 (NFR06 - Security)*

|  |  |
| --- | --- |
| **Req. No.** | **Non-Functional Requirement** |
| **NFR06-1** | The system must only allow authorized users with correct login credentials to access the system. |

In the above Table 11, non-functional requirements of the software’s security are mentioned.

#### NFR-07: Extensibility

*Table 12 (NFR07 - Extensibility)*

|  |  |
| --- | --- |
| **Req. No.** | **Non-Functional Requirement** |
| **NFR07-1** | The system must be built in a modular form so that any extension can be easily integrated |

In the above Table 12, non-functional requirements of the software’s extensibility are mentioned

## 2.5 Use Case Description:

*Table 13 (Use case - 1)*

|  |  |
| --- | --- |
| **Use Case ID:** | 1 |
| **Use Case Name:** | User Login |
| **Actors:** | User |
| **Description:** | The User wants to access the Application. |
| **Pre-Condition:** | The User should know login credentials. |
| **Post-Condition:** | The User will be provided access to the system. |
| **Normal Flow of Events:** | 1. The user opens the web application. 2. The user adds his/her email and password. 3. The user receives access to application. |
| **Alternatives Flow:** | 1. The admin enters incorrect login credentials. 2. Access to the system is denied |
| **Exceptions:** | None |

Table 13 shows the **Use Case ID: 1**. In this module, the user enters the login credentials. The software compares the information with the information stored in the database and if the information is correct, the user is provided access to the system else denied the access with appropriate message.

**Text to Speech:**

*Table 14 (Use case - 2)*

|  |  |
| --- | --- |
| **Use Case ID:** | 2 |
| **Use Case Name:** | Text to Speech |
| **Actors:** | Application Admin, User |
| **Description:** | Once user login into application, then can use the feature of text to speech, which means user provide its text in text box and select language of text, according to it can select targeted person voice from provided options. After Clicking on convert button, input text will be processed and output will be shown within 2 to 3 minute. If any error or text not processed, then appropriate message will be displayed.  Application Admin have the control of updating feature like adding new targeted person voice and more. |
| **Pre-Condition:** | The user need to login first to access this feature.  Text provided should be in proper English which means no grammatical mistakes as well as no ambiguity. |
| **Post-Condition:** | The user provided with generated speech output.  If any error then appropriate message displayed to user. |
| **Normal Flow of Events:** | 1. The user should be login into the Application. 2. User provide its text in text box for speech generation. 3. Select the target person voice from given options. 4. Output audio display to the user in web Application |
| **Alternatives Flow:** | 1. If input text not processed or any error, then appropriate message display to user. |
| **Exceptions:** | None |

Table 14 shows the **Use Case ID: 2**. In this module, the model is Text to speech feature, which takes text as input and provide speech as output.

**Voice Cloning:**

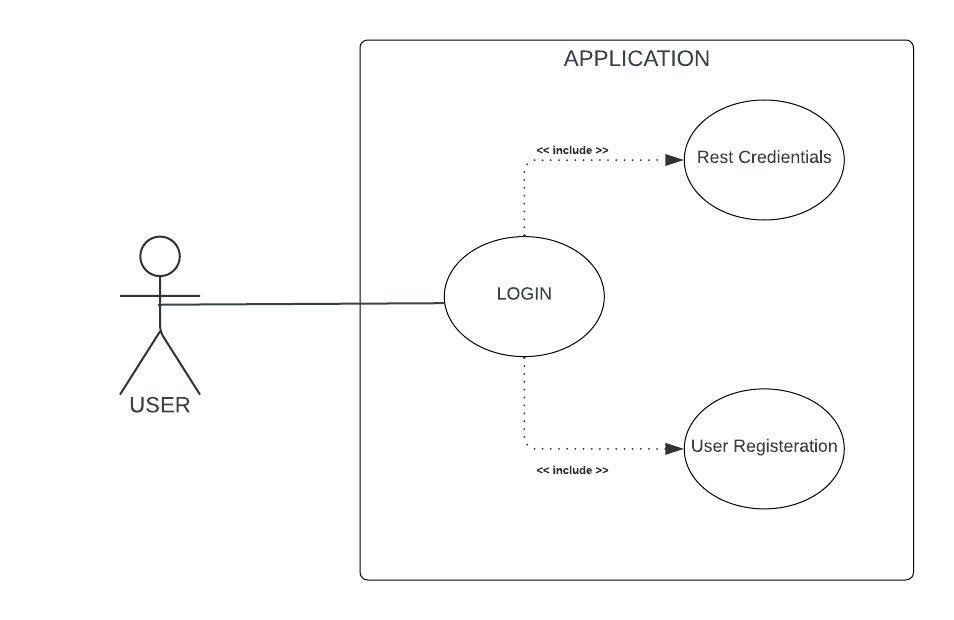
*Table 15 (Use case - 3)*

|  |  |
| --- | --- |
| **Use Case ID:** | 3 |
| **Use Case Name:** | Voice Cloning (speech to speech) |
| **Actors:** | Application Admin, User |
| **Description:** | Once user login into application, then can use the feature of voice cloning, which means user provide audio file using microphone or import file from device and select language of input audio, according to it can select targeted person voice from provided options. After Clicking on convert button, input audio will be processed and output will be shown within 2 to 3 minute. If any error or audio not processed, then appropriate message will be displayed as output.  Application Admin have the control of updating feature like adding new targeted person voice, enhancing existing features and more. |
| **Pre-Condition:** | The user need to login first to access this feature.  Input audio provided should be clear, in English and no background noise as it might will effect generated output. |
| **Post-Condition:** | The user provided with generated speech output.  If any error then appropriate message displayed to user. |
| **Normal Flow of Events:** | 1. The user should be login into the Application.   2. User provide its audio as input for speech generation.  3. User should select the language of text (English).   1. Select the target person voice from given options. 2. Output audio display to the user in web Application |
| **Alternatives Flow:** | 1. If input audio not processed or any error, then appropriate message display to user. |
| **Exceptions:** | None |

Table 15 shows the **Use Case ID: 3**. In this module, the model is voice cloning feature, which takes input as audio and provide speech as output.

**Use Case Diagram:**

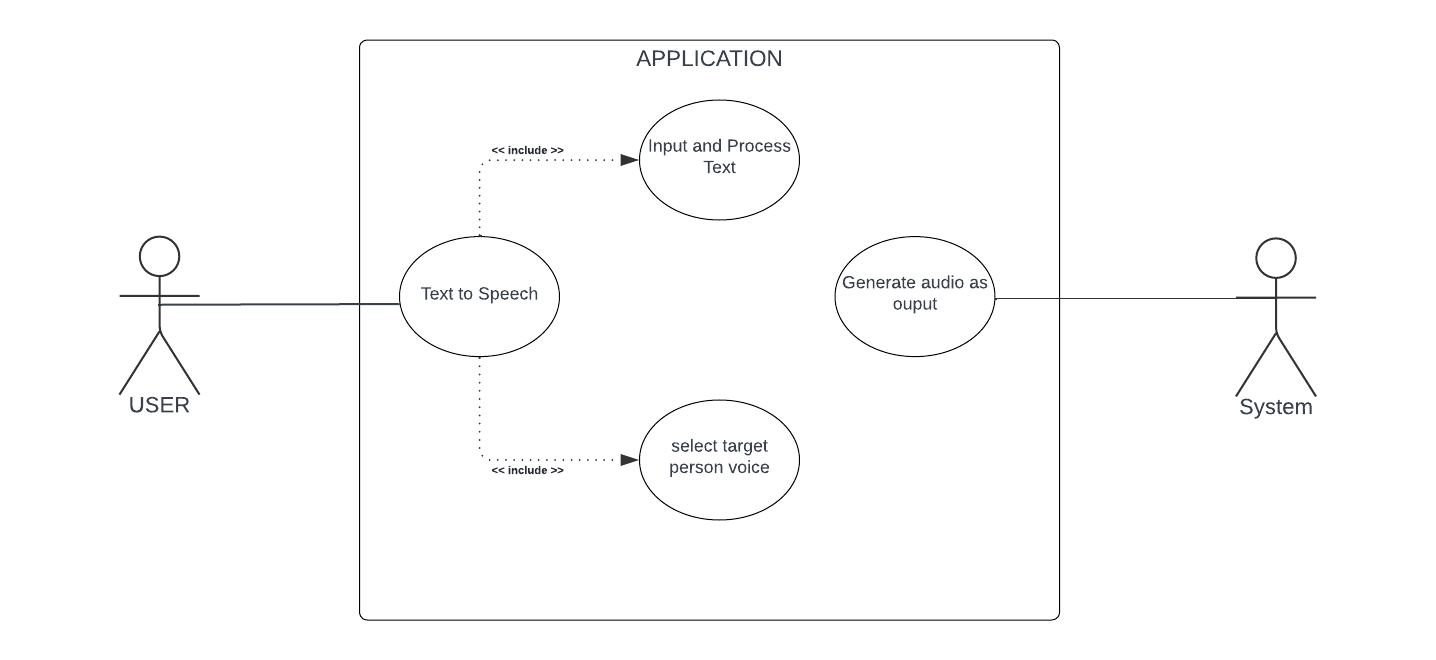
**Use Case ID-01:**



*Figure 01 (Login))*

Figure 01 above presents the description for Use Case 1, which is User Login. The actor in this use case is User who provides the right credentials to login into the system or the access is denied.

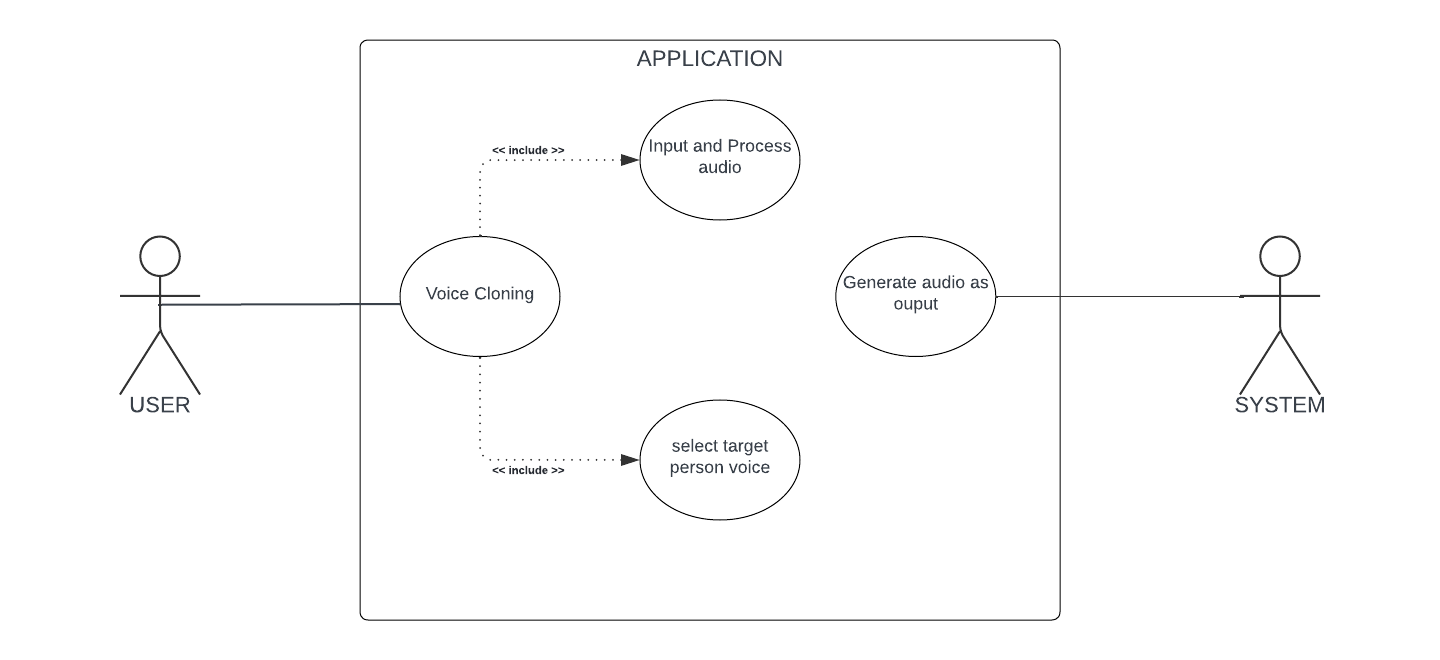
**Use Case ID-02:**



*Figure 02(text to speech)*

Figure 02 above presents the description for **Use Case 02**, which is **Text to Speech**. The actor in this use case is User can provide the input in form of text and have a generate output in form of audio.

**Use Case ID-03:**



*Figure 03 (voice cloning)*

Figure 03 above presents the description for **Use Case 03**, which is **Voice Cloning**. The actor in this use case is User can provide the input in form of audio and have a generate output in form of audio.

## 2.6 Software Development Life Cycle:

According to our Application **Deepfakes text to speech**, at initial stage, our application is considering some functionality which can be enhanced in future or can add some other Features later. Its mean our Project need to be launch first with fewer features and keep on updating with time. So, we are considering **Agile Development Methodology**. Figure of **Agile model** is given below



As Agile is a very popular development methodology used by many big tech Companies for their projects. It’s a very **fast development** life Cycle methodology which means first with one product feature like in our case text to speech having only one target person voice, we can deploy our application and check its working using **alpha-beta testing.** After reviewing **feedback**, can made enhancement in existing feature or add new functionality. Each product increment in Application of fixed length are called **Sprints** in Agile methodology. Some advantages of agile are mentioned below for our case

**Advantages:**

* **Flexibility and Adaptability:** Agile allows for changes to be made even late in the development process. As it can help us if we want to change our requirement little later.
* **Iterative Development:** Agile breaks down the project into small, manageable iterations or **sprints**. Which help us in developing the Application first, for one feature and do its alpha-beta testing.
* **Adaptive Planning:** Agile projects embrace change. Our team can regularly reviews and adapts plans based on feedback.
* **Continuous Improvement:** The Agile methodology encourages a culture of continuous improvement. Our Teams can regularly reflect on their processes and outcomes, making adjustments to enhance efficiency and effectiveness.

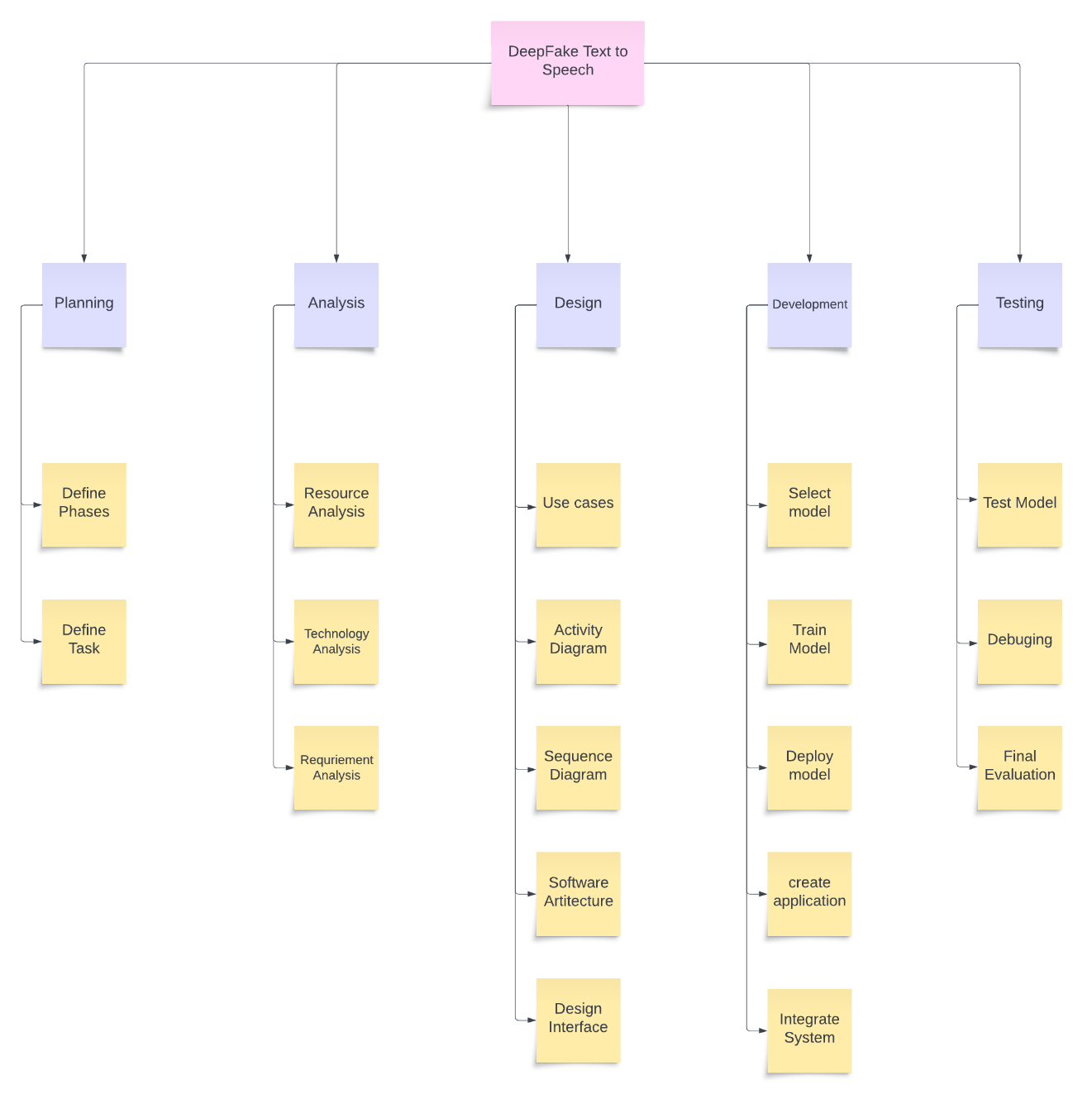
**Major Future Upgradation in Project:**

* **Adding new target voices:** Adding new target person voices for text to speech and voice cloning features, which will offer user with more options.
* **Adding more languages:** Our application, both features text to speech and voice cloning will be offered in more languages, our major focus will be on **Urdu**.
* **Deepfakes detection system:** After completing major functionalities, we will be considering to work on developing AI based Deepfakes detection system and integrating with our project.

Chapter 3: System Design

# System Design:

## Work Breakdown Structure: (WBS)

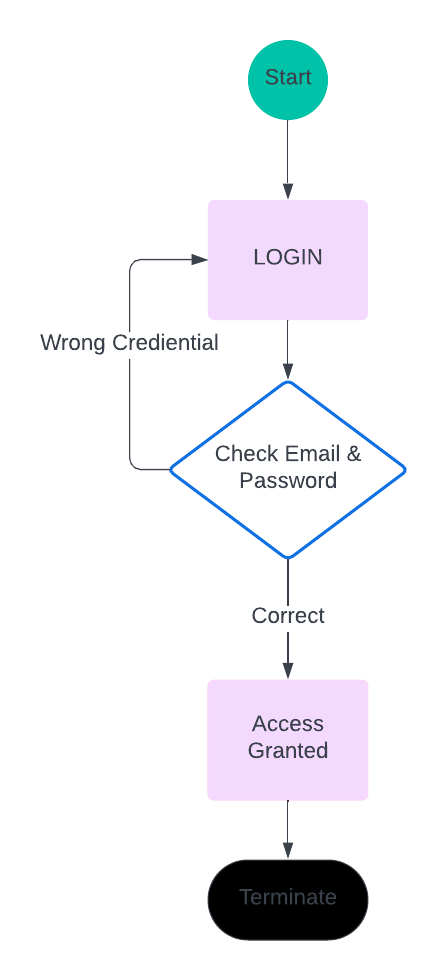


*Figure 04 (Work Breakdown Structure)*

The Figure 04 above presents the entire **Work Breakdown Structure** (WBS) of the Deepfakes text to speech, including all its phases for Planning, Analysis, Designing, Development and finally, the System Testing.

## 3.2 Activity Diagram:

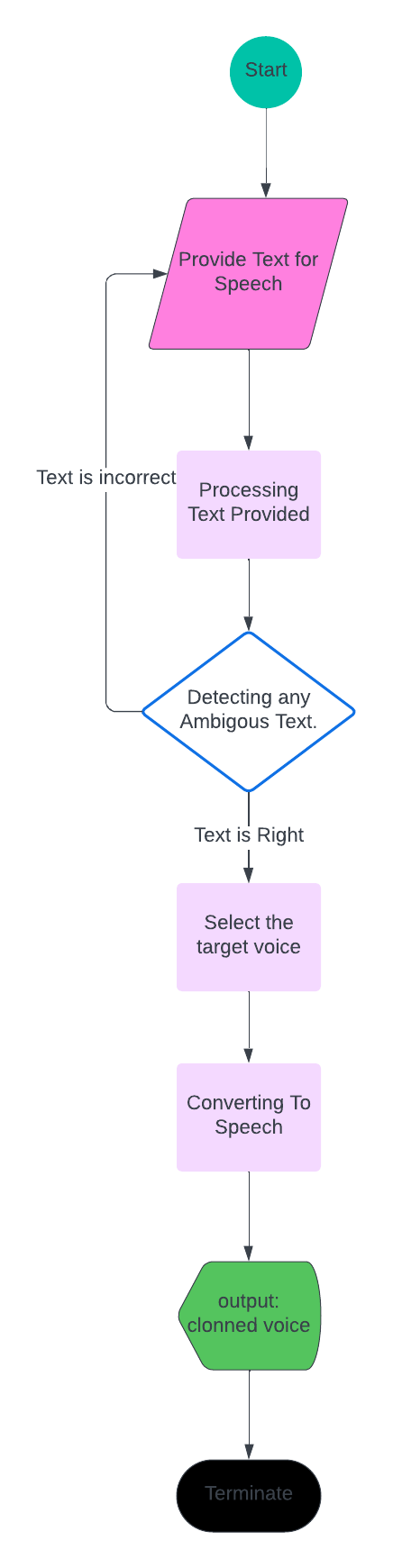
### 3.2.1 Login:



*Figure 05 (Activity - 1)*

The Figure 05 above presents the activity sequence for system **login**.

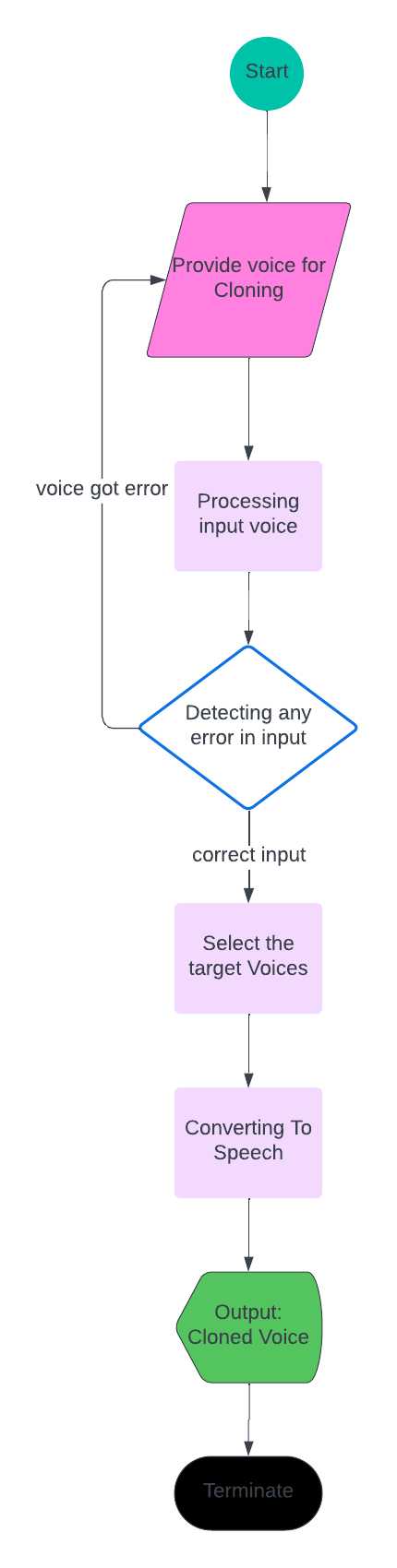
### 3.2.2 Text to Speech:



*Figure 06 (Activity - 2)*

The Figure 06 above presents the activity sequence for **Text to Speech**.

### 3.2.3 Voice Cloning:

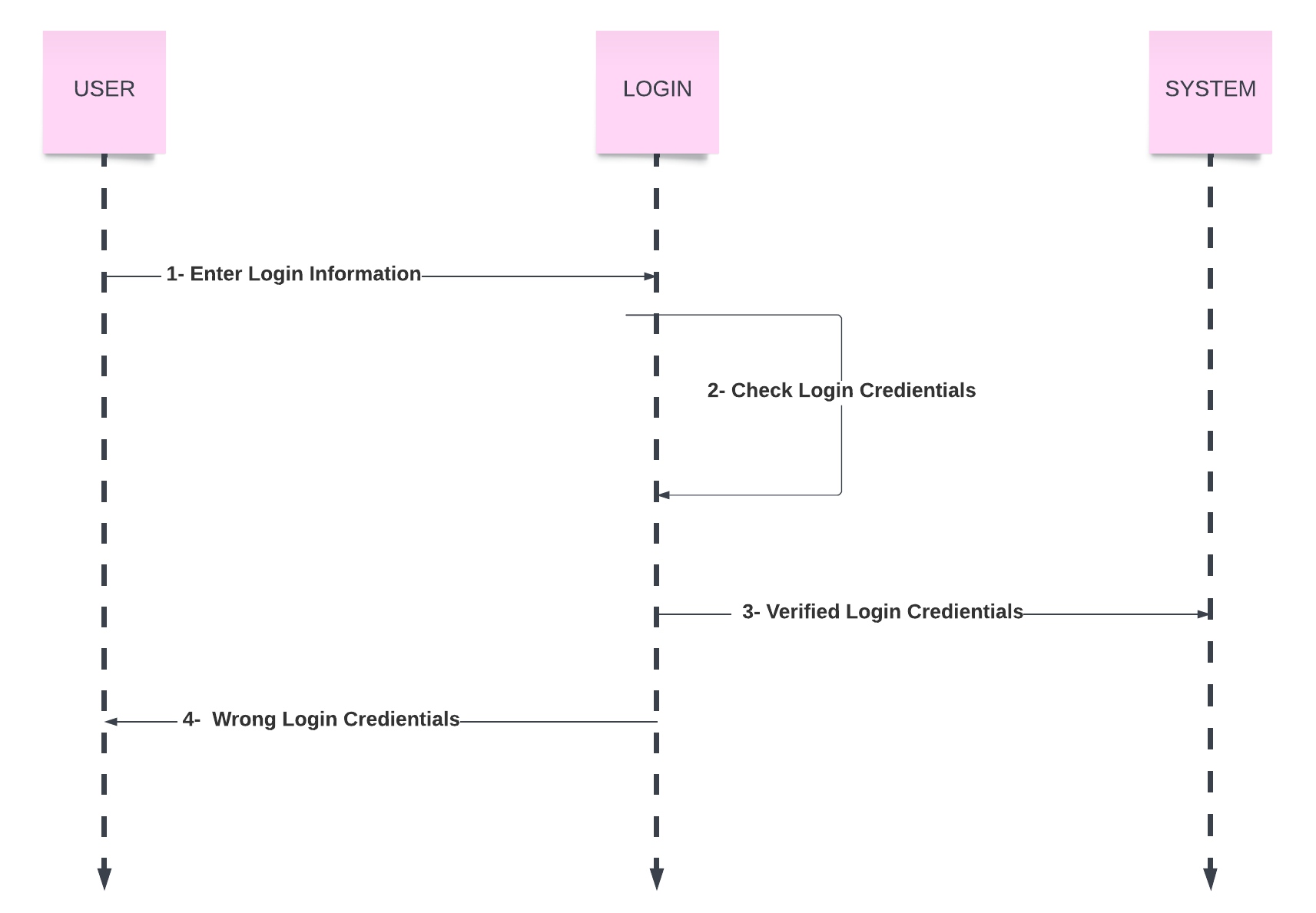


*Figure 07 (Activity - 3)*

The Figure 07 above presents the activity sequence for **Voice Cloning**.

## Sequence Diagram:

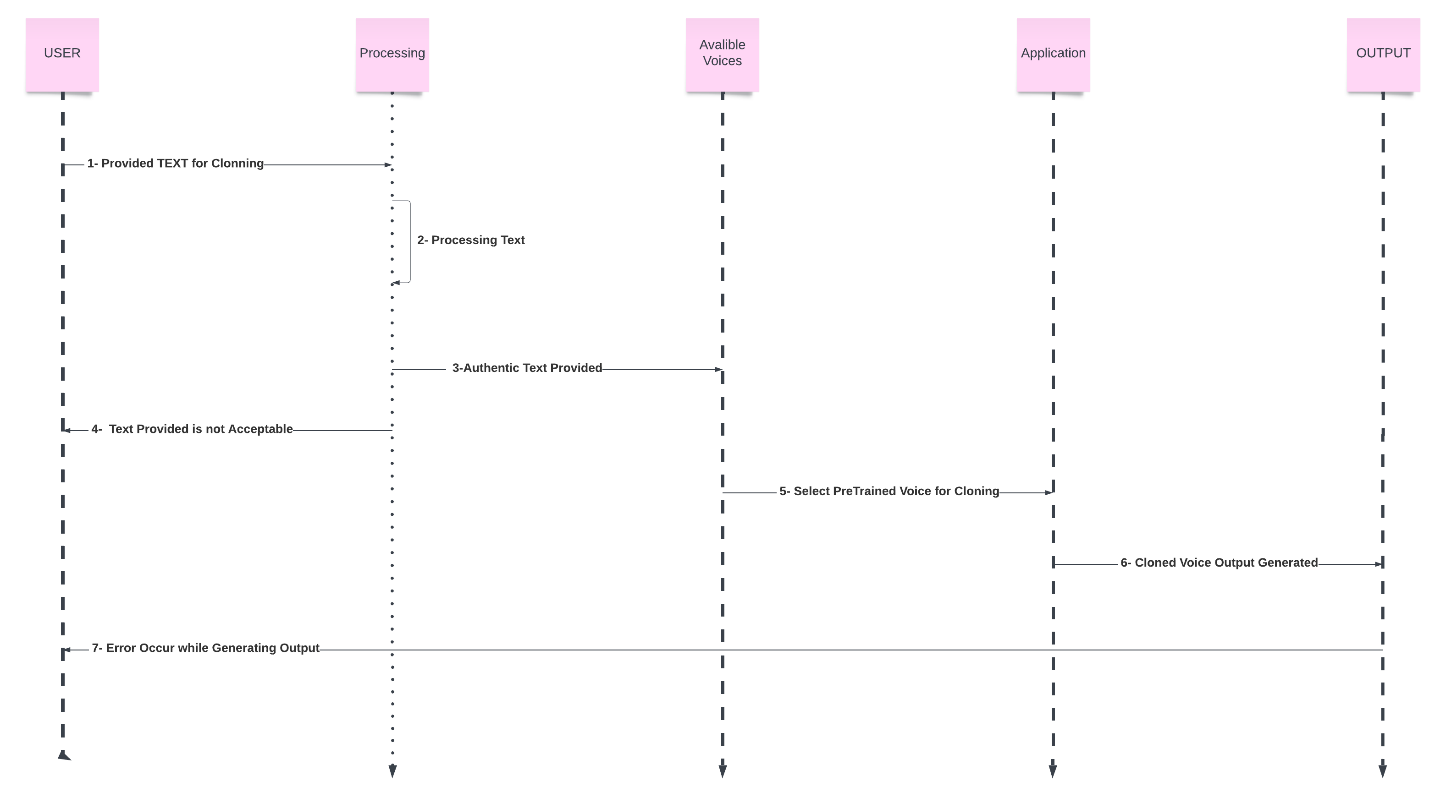
### Login:

****

*Figure 08 (Sequence- 1)*

The Figure 08 above describes the sequence for User **Login** into the Deep cloning application.

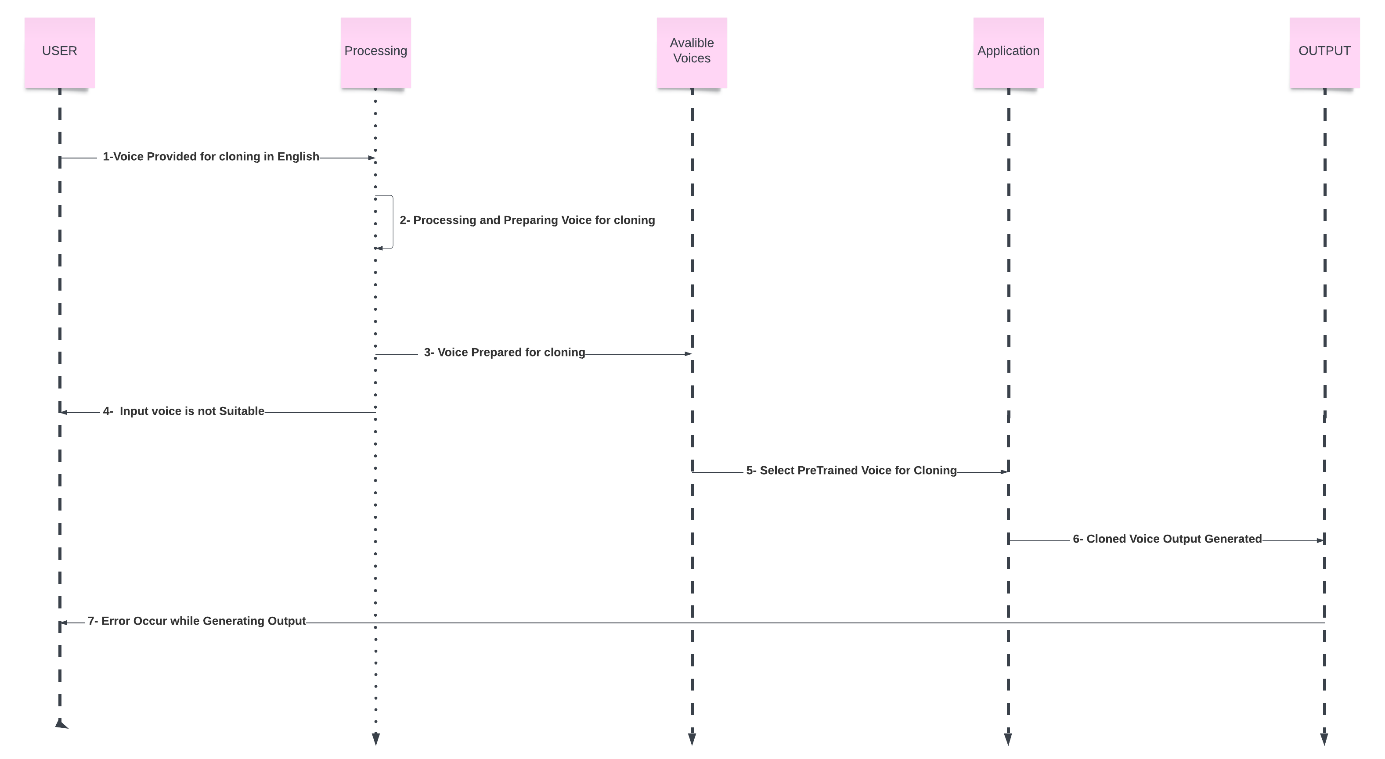
### Text to Speech

****

*Figure 09(Sequence - 2)*

Figure 09 above describes the sequence for the **Text to Speech**.

### Voice Cloning

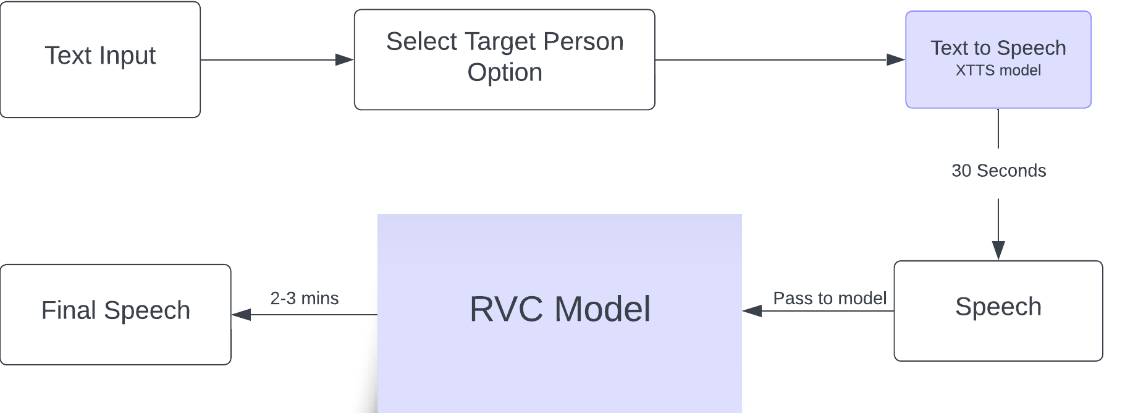
****

*Figure 10 (Sequence - 3)*

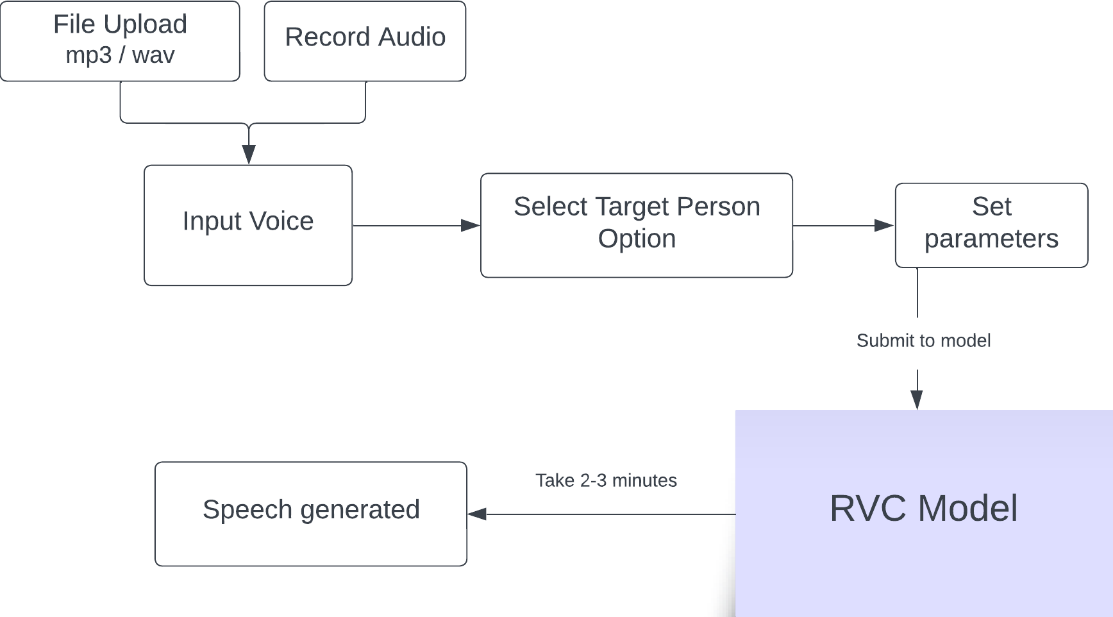
Figure 10 above describes the sequence for the **Voice Cloning**

## Software Architecture:

**Text-to-Speech:**



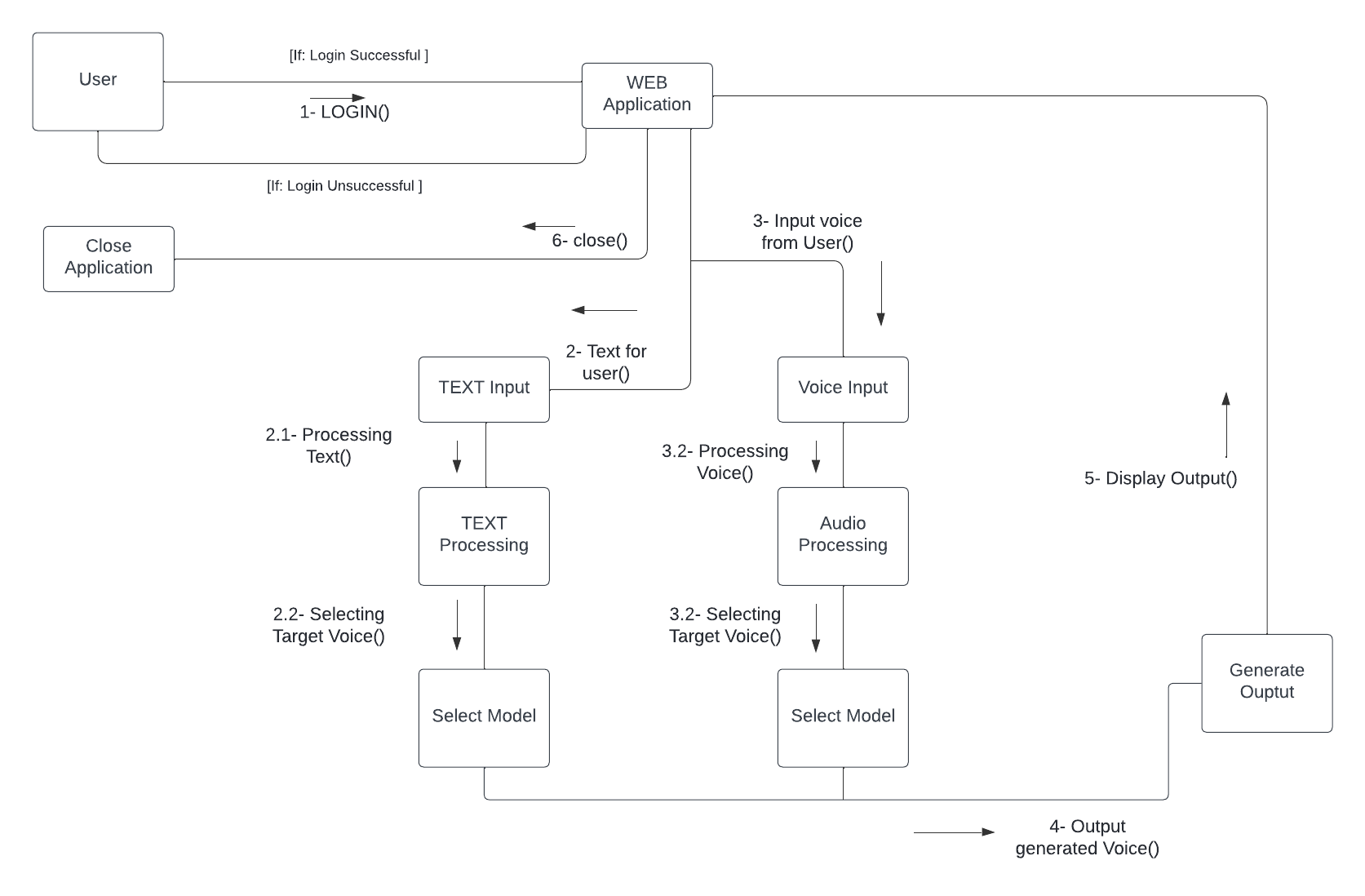
**Voice Cloning:**



*Figure 11 (Architecture)*

The Figure 11 above presents the System Architecture for the Deep Cloning (Deepfake text to speech) Application.

## 3.4 Collaboration Diagram



*Figure 12 (Collaboration)*

The Figure 12 above presents the **Collaboration sequence** of the entire system of Deep cloning web application.

Chapter 4: System Testing

# System Testing

## Test Cases

*Table 16 (Test Case - 1)*

|  |  |
| --- | --- |
| **Test case ID: TC-01** | |
| Application Name | Deepfakes Text to Speech |
| Use case(s) | Login |
| Input summary | User will be asked to provide their login details |
| Pre-condition | The application should be turned on |
| Post-condition | User will be navigated to the home screen |
| Output Summary  *IF SUCCESS:*  Use is provided access  *ELSE:*  Error message will be displayed about wrong credentials and user will be asked to enter again | |

In the above Table 16, a test case regarding login to user.

.

*Table 17 (Test Case - 2)*

|  |  |
| --- | --- |
| **Test case ID: TC-02** | |
| Application Name | Deepfakes Text to Speech |
| Use case(s) | Voice cloning |
| Input summary | Input voice which we want clone |
| Output | Model train on the target voice |
| Output Summary  *IF SUCCESS:*  When model train it can give voices | |

In the above Table 17, a test case regarding voice cloning.

*Table 18 (Test Case - 3)*

|  |  |
| --- | --- |
| **Test case ID: TC-03** | |
| Application Name | Deepfakes Text to Speech |
| Use case(s) | Text to Speech |
| Input summary | Giving text for speech |
| Pre-condition | Select the voice |
| Post-condition | Speech is generated for target input |
| Output Summary  *IF SUCCESS:*  Output speech is generated against input text. | |

In the above Table 18, a test case regarding to text to speech.

*Table 19(Test Case - 4)*

|  |  |
| --- | --- |
| **Test case ID: TC-07** | |
| Application Name | Deepfakes Text to Speech |
| Use case(s) | Create Account |
| Input summary | User should open create account screen |
| Pre-condition |  |
| Post-condition | New profile will be created |
| Output Summary  *IF SUCCESS:*  New profile will be created and saved in the database | |

In the above Table 19, a test case regarding new user profile is described.

## Unit Testing

In this phase, we tested our project’s individual components that perform unit tasks and are part of the whole workflow of the project.

The limited bugs that were identified and fixed were:

## Acceptance Testing

In this phase, we checked our project to see whether it fulfils our business requirements.

In our case, we check the following functionalities of the software:

* Login
* Voice cloning
* Text to speech

Our software fulfilled the above listed requirements and is ready to be deployed.

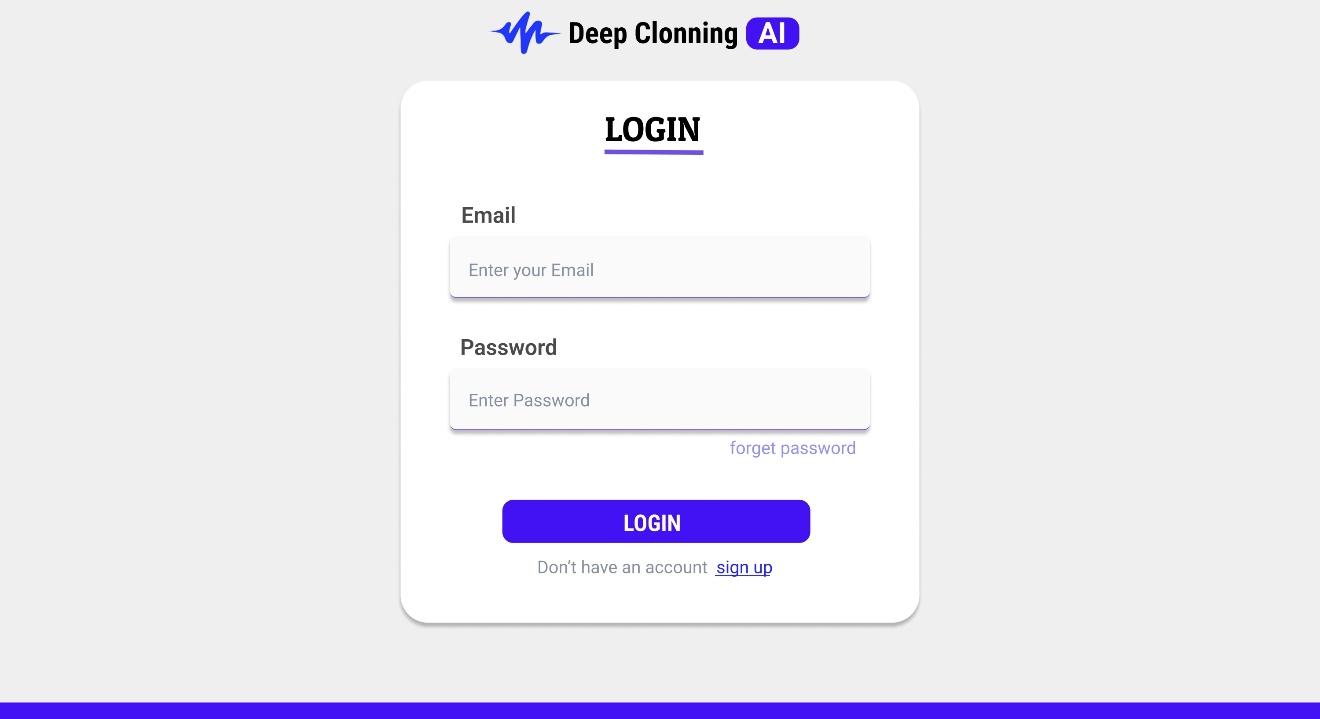
Chapter 5: Application Front-End

# Application Frontend

This section contains screenshots of the various screens of Deep Cloning Web application Interface.

## Web Interface

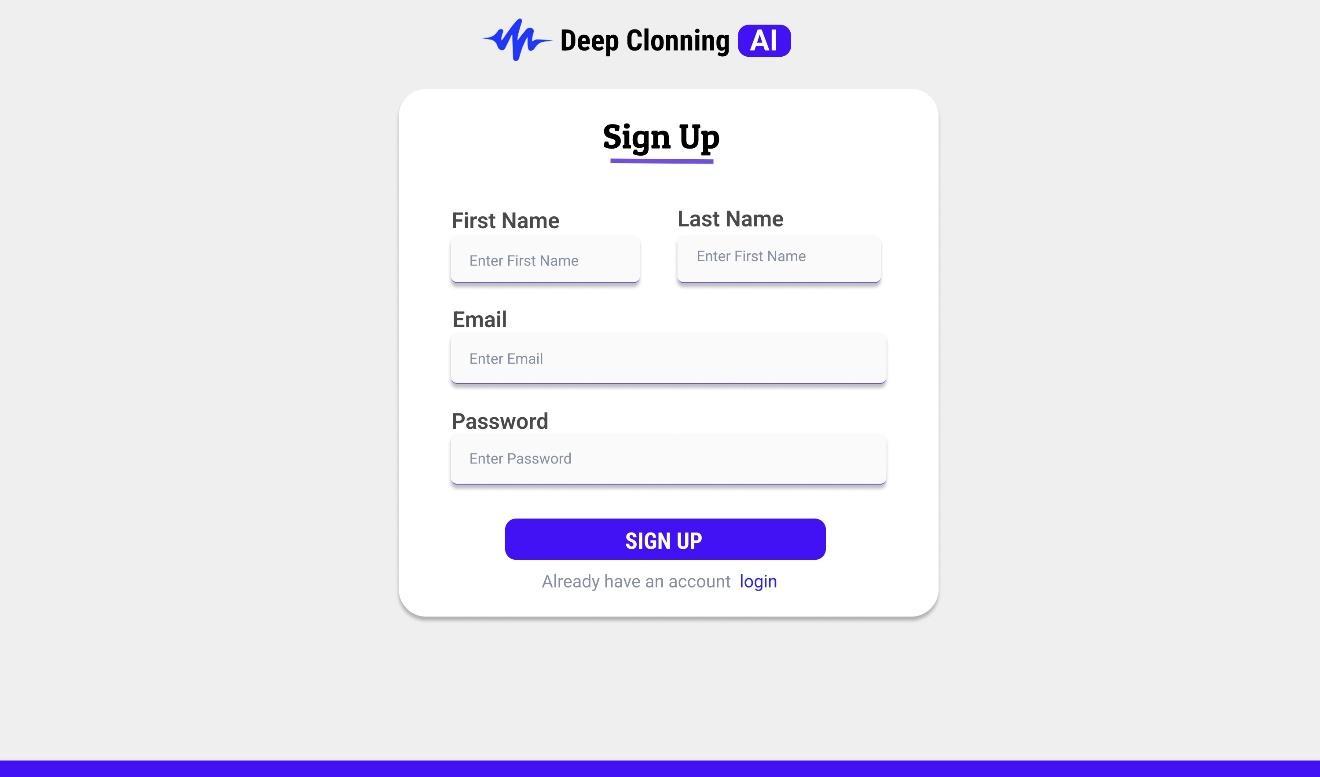
### Login Screen



*Figure 13 (Login)*

The above figure 13 tells that login screen of the app.

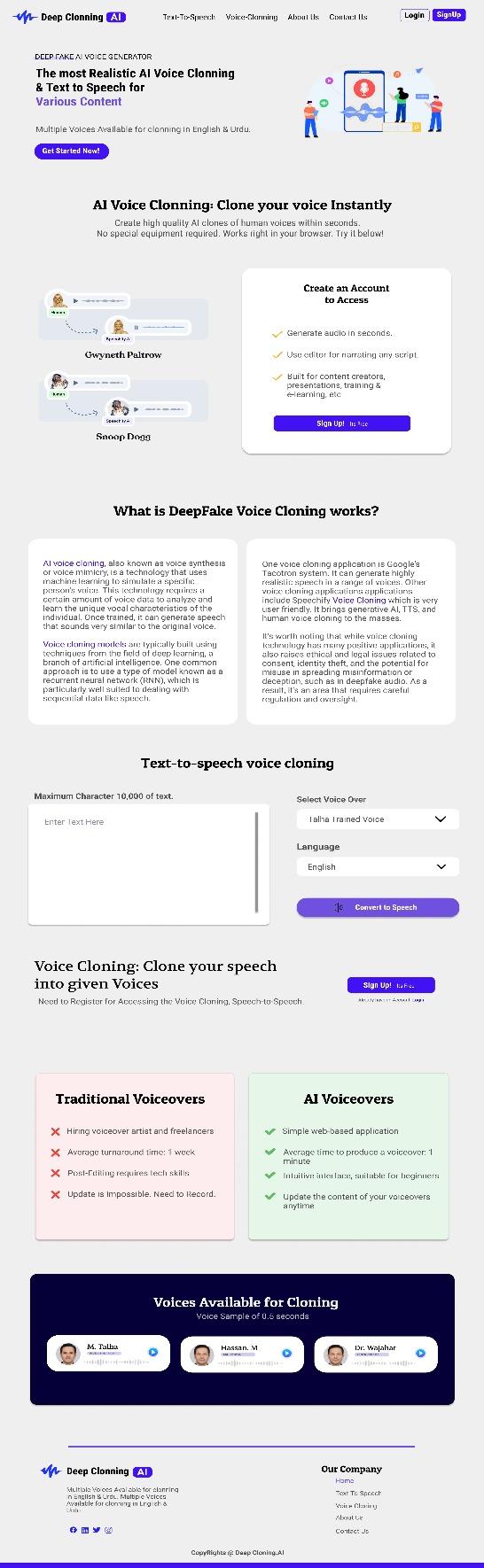
### Sign Up Screen



*Figure14 (Signup)*

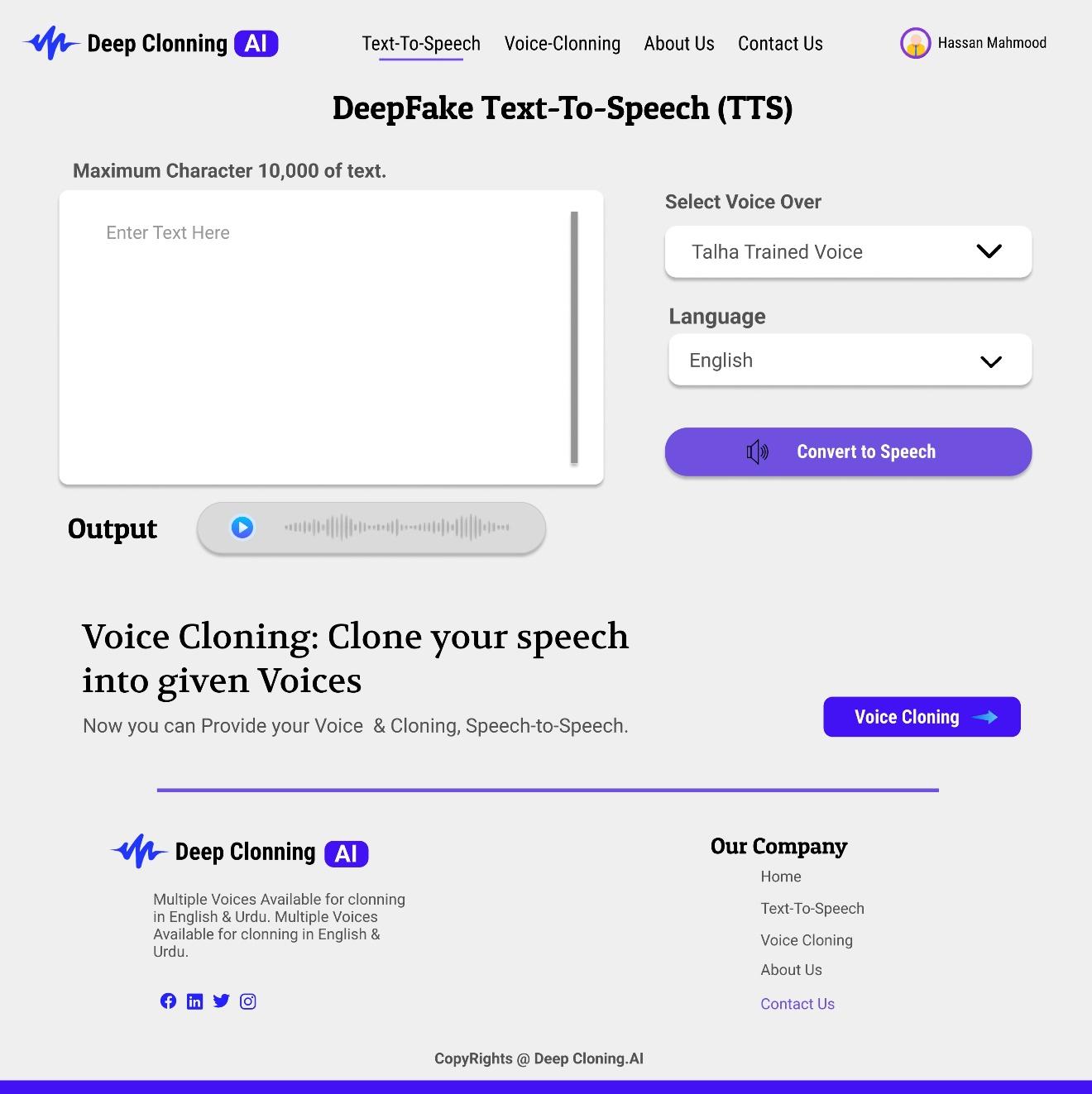
The above figure 14 is regarding about signup page of the app.

### Home Screen



*Figure15 (Home Screen)*

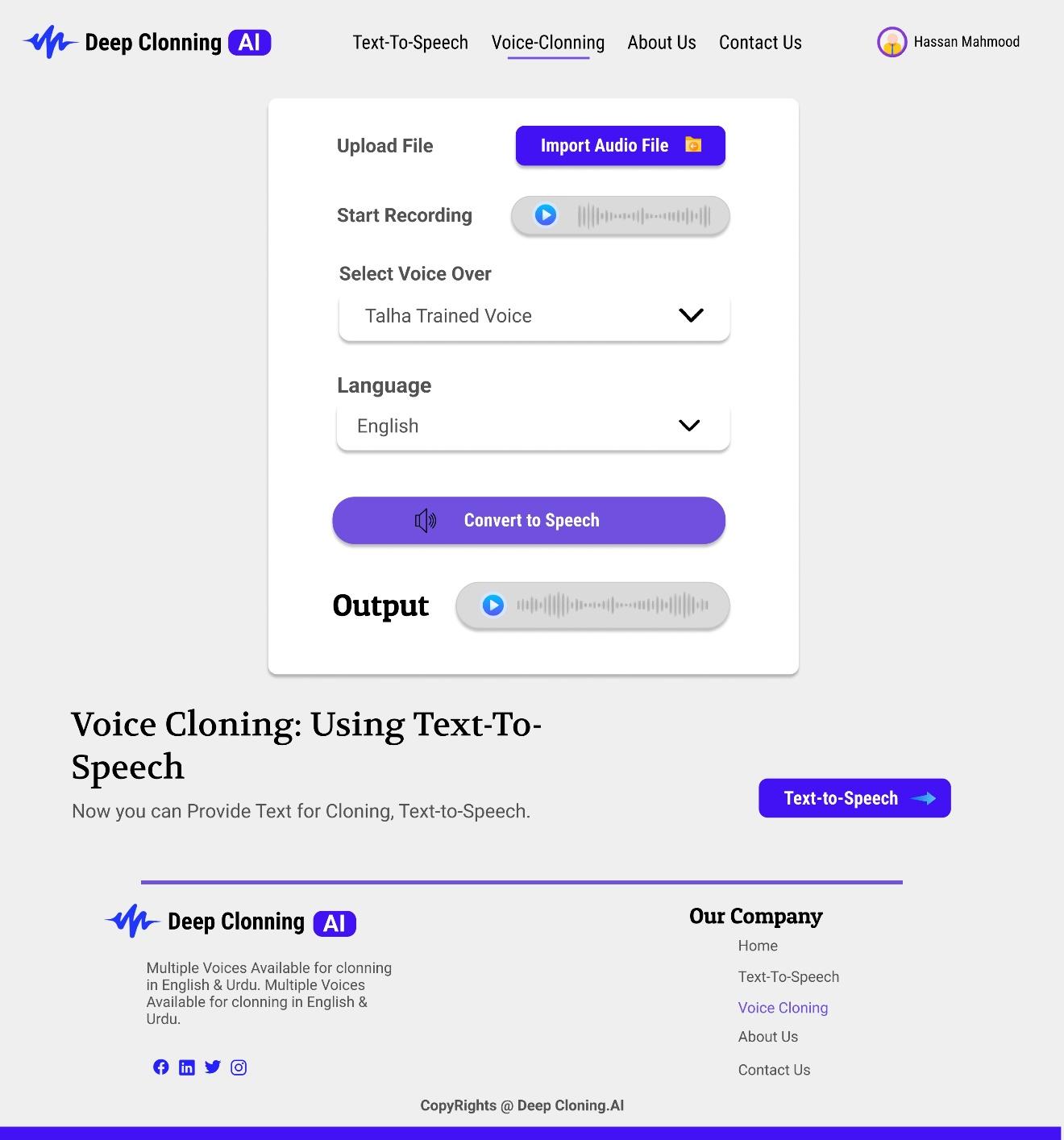
### Text to Speech



*Figure 16 (Text to Speech)*

The above figure 16 is about converting text to speech.

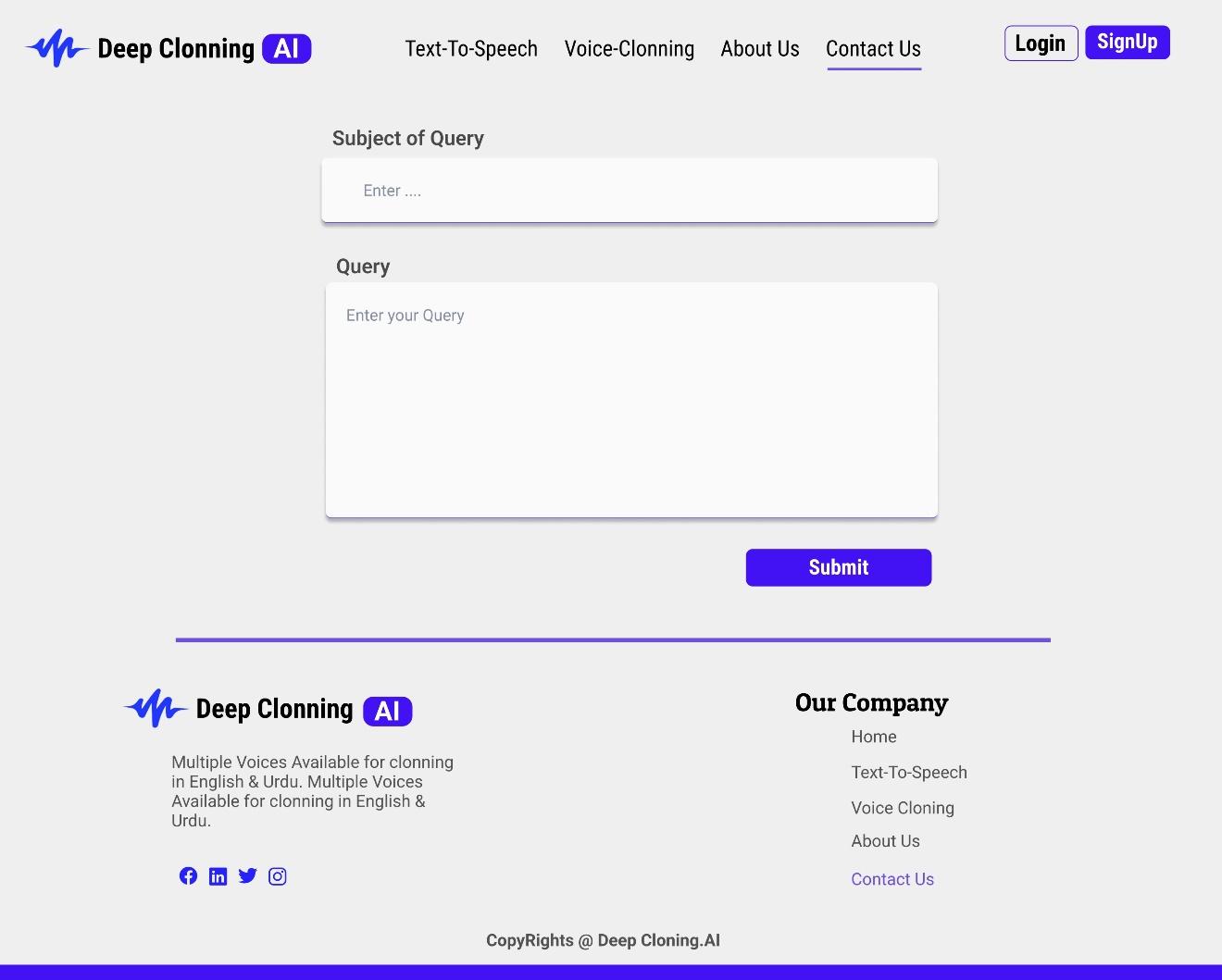
### Voice Cloning



*Figure 17 (Voice Cloning)*

The above figure 17 is about voice cloning of the user.

### Contact Us



*Figure 18 (Contact Us)*

The above figure 18 is about the contact us page of the app.

Chapter 6: Conclusion

# Conclusion

In summary, our project at Expert System Solution focused on diving into the world of Deepfakes technology, shedding light on its development and impact on society. Specifically, we honed in on crafting Deepfakes text-to-speech synthesis, centre our efforts on a single-person dataset. Real-world instances have demonstrated that Deepfakes are now widespread, affecting various sectors, notably in news media, where they present challenges tied to misinformation and public manipulation.

Throughout the project, we explored the capabilities of deep learning tools like GANs, recognizing the double-edged nature of technology that can be both creatively empowering and potentially harmful. The growing interest in Deepfakes tech is apparent, seen in the increased number of articles and discussions on the subject.

Looking forward, our vision is to contribute to the field by creating tools for detecting and preventing Deepfakes. We also plan to broaden our project's scope, encompassing multiple persons and delving into the realm of personal voice Deepfakes. Our ultimate goal is to tackle the challenges posed by Deepfakes tech, striving to foster responsible and ethical applications of artificial intelligence in the ever-changing landscape of information technology.

## Problem faced

Certainly, the field of deepfake text-to-speech (TTS) has its share of challenges and issues. Here are some common problems faced in deepfake TTS, particularly with models like Suno Bark and RVC:

1. **Noise in Output:**

- The generated audio often contains noise, artifacts, or distortions that reduce the overall quality and naturalness of the voice. Cleaning the audio post-generation becomes necessary, requiring additional processing steps.

2. **Duration Limitations**:

- Models like Suno Bark have limitations on the duration of the generated output (e.g., no output beyond 2-3 minutes). This can be restrictive for users who need longer audio segments.

3. **Quality Dependency**:

The accuracy and naturalness of the output heavily depend on the quality of the input voice. If the input is not clean or of poor quality, the generated output may lack clarity and coherence. We uses lots of models like **Tortoise\_TTS**, **Suno-Bark**, and **XTTS** just to check quality of generated output.

4. **Limited Expressiveness**:

- Some deepfake TTS models struggle to capture and express a wide range of emotions, intonations, or nuances present in human speech. This limitation may result in monotonous or less expressive voice outputs. Moreover, no model available currently which can clone accent of target person completely.

**5. Resource Intensiveness:**

Training and using deepfake TTS models can be computationally intensive, requiring significant computing resources. This can be a barrier for smaller projects or individuals with limited access to high-performance hardware.

**6.2 Project Summary:**

Our project specifically focuses on developing Deepfakes text-to-speech synthesis using a single-person dataset, with selected datasets playing a crucial role in training our models. The ultimate goal is to address the challenges posed by Deepfakes technology, and developing a System which should be nearly accurate in cloning the targeted person voice. Yes, there are Challenges in Achieving this level but still satisfactory results achieved after testing lots of approaches and models.

Chapter 7: References

# References

|  |  |
| --- | --- |
| [1] | J. N. B. Jeffrey T. Hancock, “The Social impact of Deefakes.,” https://doi.org/10.1089/cyber.2021.29208.jth, 2021. |
| [2] | J. Gadgilwa, K. Rahangdale, O. Jaiswal, P. Asare, P. Adekar and P. L. Bitla, “Exploring Deepfakes - Creation Techniques, Detection Strategies, and Emerging Challenges: A Survey,” 2020. [Online]. Available: https://doi.org/10.22214/ijraset.2023.49681. [Accessed 2023]. |
| [3] | O. TV and K. P. Bhavsar, “Odisha gets 1st AI-generated news anchor 'Lisa' | Watch,” 2023. [Online]. Available: https://www.livemint.com/news/india/odisha-gets-1st-ai-generated-news-anchor-lisa-watch-11688969170338.html. [Accessed 2023]. |
| [4] | S. Burgess, “Ukraine war: Deepfake video of Zelenskyy telling Ukrainians to 'lay down arms' debunked,” Sky News, 2022. [Online]. Available: https://news.sky.com/story/ukraine-war-deepfake-video-of-zelenskyy-telling-ukrainians-to-lay-down-arms-debunked-12567789. |
| [5] | Z. Khanjani, G. Watson and V. P. Janeja, “Audio DeepFakes: A Survey,” 09 January 2023. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fdata.2022.1001063/full. [Accessed 2023]. |
| [6] | “Retrieval-based-Voice-Conversion-WebUI,” [Online]. Available: https://github.com/RVC-Project/Retrieval-based-Voice-Conversion-WebUI. [Accessed 2023]. |
| [7] | Kalomaze, “RVC v2 AI Cover Guide.,” [Online]. Available: https://docs.google.com/document/d/13\_l1bd1Osgz7qlAZn-zhklCbHpVRk6bYOuAuB78qmsE/edit#heading=h.qjrl2d41vtmt. [Accessed 2023]. |
| [8] | Revoice.AI, “Revoicer.AI,” Audio voice Cloning, [Online]. Available: www.revoicer.AI. [Accessed 2023]. |
| [9] | Speechify, “Speecify.AI,” Text to Speech and Voice Cloning, [Online]. Available: www.Speechify.AI. |