Arab International University

Faculty of Informatics and Communication Engineering

Senior Project Report on

AI Voice Dubbing

Submitted to

Department of Informatics Engineering

in partial fulfillment of the requirement for the Degree of Bachelor in

Informatics Engineering

Submitted by:

Mouna Al-Tahhan **Maria Ammash**

Laith Al Mouzain

Mhd Safouh Ghazal **Judy Sweed**

Under the Supervision of:

Dr.Tarek Barhoum **Eng.Mohammad AlMasri**

July 2024

July 2024

Faculty of Informatics & Communication Engineering

CERTIFICATE OF APPROVAL

The undersigned certify that they have read and recommended to the Department of Informatics Engineering for acceptance, a project report entitled “AI Voice Dubbing” Submitted by: Mouna Al-Tahhan, Maria Ammash, Judy Sweed, Mhd Safouh Ghazal and Laith Al Mouzain in partial fulfillment for the degree of Bachelor of Engineering in Informatics.

Supervisor

Dr. Tarek Barhoum

Head of Department

Dr. Said Desouki

Arab International University

The Arab International University (AIU) is a private Syrian university established in 2005. Its academic plans and the documents issued by it are approved and certified by the Ministry of Higher Education in the Syrian Arab Republic.

The university works to achieve the following goals:

- Preparing a distinguished generation of university graduates who are able to meet and advance the specific needs of society.
- Contributing to theoretical and applied scientific research that serves the purposes of national development. Work is being done to urge professors and academic staff to scientific research and participate in conferences and seminars that organize research.
- Achieving partnership with prestigious Arab and foreign universities with the aim of continuous development and modernization of academic work and conducting joint scientific research.
- Attracting distinguished academic and research competencies by providing the appropriate environment for their work.

The Arab International University is one of the first Syrian universities that have been established and inaugurated. It has been able to attract distinguished educational, research and administrative competencies, to create an integrated edifice from the academic, organizational and administrative aspects. It was able to graduate cadres of distinguished innovators by providing an educational environment based on unique qualitative and material ingredients, including:

- Modern and advanced study plans based on the credit hour system.
- Carefully selected educational cadres.
- Modern scientific laboratories and a laboratory for electronic libraries.
- Physical and moral incentives for students.
- Application of interactive teaching methods.
- Academic and educational guidance and counseling.
- A wide range of scientific cooperation agreements with local, regional and international universities of reputable reputation.
- Multiple agreements and memoranda of understanding with many civil society organizations.
- A proper campus with all the facilities of science, sports and entertainment, which we encourage you to visit and learn about their features.
- Student activities and clubs of all kinds: athletic, cultural, scientific and social.

The Arab International University life years are a time to invest in a student's future.

The knowledge and experience that students acquire in the lecture hall and laboratories will help them in developing themselves. They will provide them with reasons for success in the chosen specialty. The student activities will help in expanding students' horizons. The activities of training, clubs and sports will enable students to develop their talents, and may even help in discovering new talents.

Students could invest time, mind and spirit in our university in order to reap the benefits of work, and the time devoted in the coming years. We will be by our students in every step of their way.

الإهداءات

إلى أساس نجاح هذا المشروع وخروجه بأفضل صورة ممكنة
تتسابق الكلمات وتتزاحم العبارات توفيك حقك من الشكر
إلى من رافقنا في مسيرتنا لإنجاز هذا البحث واستمرارها يدفعنا إلى الأمام

الدكتور طارق برهوم

إلى العظيم الذي أنار هذا الدرب
وصاغ من علومه ومعرفته منارة لا تنطفئ
الراقي الذي آمن بقدراتنا والمشرف على المشروع

المهندس محمد المصري

إلى عميد الكلية الفاضل، يُشكر على إدارته الناجحة للكلية وسعيه الدائم لتطويرها، ووجوده الدائم
للاستماع لمشاكل الطلاب وحاجاتهم

الدكتور العميد فايز كيوان

وأسمى آيات الشكر والتقدير للذين حملوا أقدس رسالة في الحياة
إلى الذين مهدوا طريق العلم والمعرفة بأفضل صورة ممكنة

أساتذتنا الأفاضل في الجامعة العربية الدولية

Acknowledgements

It gives us immense pleasure to express our deepest sense of gratitude and sincere thanks to our highly respected and esteemed guide **Dr.Tarek Barhoum** for his valuable guidance, encouragement and help for completing this work, his useful suggestions for this whole work and cooperative behavior are sincerely acknowledged.

We are also grateful to **Eng. Mohammad Al Masri** for his constant support and guidance.

At the end we would like to express our sincere thanks to all our families, friends and others who helped us directly or indirectly during this project work.

Abstract

The advent of artificial intelligence has revolutionized numerous domains, including media and entertainment. This project focuses on developing an AI Voice Dubbing application that facilitates the seamless dubbing of audio content from English to Arabic. Unlike traditional robotic voice outputs, our application ensures that the dubbed content retains the original voice characteristics, providing a natural and authentic listening experience. The system also offers users the option to select alternative output voices, including those of celebrities, or any other voice by uploading their media, providing a YouTube link, or recording audio directly. The core functionality of the application is to produce a dubbed audio file, maintaining high fidelity to the original speaker's voice tone and style, the application provides a text output of the dubbed content, which users can edit as needed to ensure accuracy and personal preference. Users can manage their dubbed projects by creating playlists and organizing them into folders with customizable names.

The application is powered by three advanced AI models: Speech-to-Text (STT), Text-to-Speech (TTS), and Voice Cloning models, which work in unison to ensure high-quality dubbing. The Speech-to-Text model accurately transcribes the spoken English content, which is then translated into Arabic. The Text-to-Speech model, enhanced by the Voice Cloning model, synthesizes the translated text into speech that mimics the original speaker's voice. This innovative approach leverages state-of-the-art AI technologies to deliver an unparalleled user experience, making it a valuable tool for content creators, educators, and media professionals seeking high-quality voice dubbing solutions.

Table of Contents

Abbreviations

AI	Artificial Intelligence
API	Application Programming Interface
ACID	Atomicity, Consistency, Isolation and Durability Theorem
MVC	Model View Control
BLoC	Business Logic Component
OAuth2	Open Authorization 2.0
LSTM	Long Short-Term Memory
BiLSTM	Bidirectional Long Short-Term Memory
NLP	Natural Language Processing
CNNs	Convolutional Neural Networks
MFCCs	Mel-frequency cepstral coefficients
GANs	Generative Adversarial Networks
ReLU	Rectified Linear Unit
DL	Deep Learning
BN	Batch Normalization
JS	JavaScript
WER	Word Error Rate
CER	Character Error Rate
STT	Speech To Text
TTS	Text To Speech

Keywords

Voice Dubbing, Audio, Text, Speech-to-Text, Text-to-Speech, Voice Cloning

Table of Figures

Introduction

In the ever-evolving landscape of artificial intelligence, advancements across various domains have achieved remarkable progress. However, amidst this growth, the demand for high-quality, natural-sounding voice dubbing solutions has become increasingly prominent. Traditional voice dubbing methods often result in robotic or synthetic outputs that detract from the listening experience, highlighting the need for more sophisticated approaches.

Today, artificial intelligence is harnessed to revolutionize the voice dubbing process through advanced AI algorithms that ensure the retention of the original speaker's voice characteristics across different data types, such as audio and video. This project focuses on developing an AI Voice Dubbing application that translates English audio content into Arabic while maintaining the natural voice quality of the original speaker. Additionally, users can select alternative output voices, including those of celebrities or any other preferred voices, by uploading media, providing a YouTube link, or recording audio directly.

Although traditional dubbing can be useful, it often lacks flexibility and quality. This application addresses these issues by offering both audio and editable text outputs, enhancing the overall user experience and ensuring the accuracy of the dubbed content. The ability to edit the text output before finalizing the dubbed audio adds a layer of customization that is critical for various applications, from educational content to professional media production.

To underscore the value of this project, consider instances where accurate and natural-sounding dubbing is essential, such as in educational videos, professional presentations, and entertainment. Providing a high-quality dubbed output not only improves the accessibility of content for Arabic-speaking audiences but also ensures that the original intent and tone of the speaker are preserved.

Addressing the need for high-quality dubbing is critical in enhancing the accessibility and reach of content, ensuring that language barriers do not impede communication. The ability to produce natural-sounding dubbed audio is paramount in maintaining the original speaker's authenticity and engagement, thereby preventing the disconnection that often accompanies synthetic voice outputs.

This technological context underscores the urgency of developing and deploying cutting-edge solutions for voice dubbing. The application not only provides users with a powerful tool for creating high-quality dubbed content but also offers tangible benefits such as enhanced user control, increased accessibility, and the ability to manage dubbed projects efficiently.

By engaging with this application, users contribute to a more connected and accessible world, fostering a sense of empowerment and collaboration in the creation and sharing of content across languages.

Chapter 1: Project Description

1.1 Background

As artificial intelligence advances rapidly, new techniques for voice dubbing have emerged, offering seamless translation and natural-sounding outputs. Traditional methods often result in robotic or synthetic voices, which detract from the user experience. Advanced AI algorithms now enable high-quality dubbing that mimics the original speaker's voice, ensuring authenticity and preserving the speaker's intent.

These AI-driven dubbing techniques provide a sophisticated way to translate audio content from one language to another while maintaining the natural characteristics of the original voice. This project leverages these advanced AI algorithms to address the challenges of achieving high-quality, natural-sounding voice dubbing from English to Arabic, enhancing user experience and ensuring the integrity of the original message.

1.2 Problem Statement

Traditional voice dubbing methods often produce robotic or synthetic outputs, detracting from authenticity and user experience. The growing demand for natural-sounding, high-quality dubbing solutions requires innovative approaches that maintain the original speaker's voice characteristics. Current challenges include the lack of flexible, user-friendly dubbing tools, high fidelity in translated audio, and ensuring dubbed content accurately reflects the original message. The objective is to develop an AI-driven application that provides seamless, natural-sounding voice dubbing from English to Arabic, enabling users to create high-quality dubbed content while preserving the original audio's integrity and intent.

1.3 Project Objective

The main objective of this project is to leverage advanced AI algorithms to provide a seamless voice dubbing solution that translates audio from English to Arabic, maintaining the original speaker's natural voice. The application also allows text editing for accuracy and efficient management of dubbed projects.

1.4 Project Scope

Enable multi-platform access for the project, ensuring the application functions on both Android and iOS. Users can register and start dubbing audio content from English to Arabic with ease. The application allows users to edit text output for accuracy and manage their dubbed projects efficiently, providing a seamless and user-friendly experience.

1.5 Project Features

The project aims to develop an artificial intelligence mobile application that offers several main features:

I. **Voice Dubbing**: Users can upload audio or video content in English and receive a dubbed audio output in Arabic, maintaining the original speaker's voice characteristics.

II. **Voice Selection**: Users can select alternative output voices, including celebrity voices

III. **Text Output and Editing**: Alongside the dubbed audio, the application provides a text output of the translated content, which users can edit for accuracy and personal

preference.

IV. **Project Management**: Users can create playlists and organize their dubbed projects into folders with customizable names, making it easy to manage multiple dubbing projects.

V. **User-Friendly Interface**: The application is designed to be intuitive and easy to use, ensuring a seamless experience for users on both Android and iOS platforms.

1.6 Project Feasibility

To determine the feasibility of the project, a comprehensive analysis of various elements is required. The project primarily relies on advanced AI algorithms for voice dubbing from English to Arabic. The key factors to consider include the development and integration of AI models, ensuring high-quality dubbing outputs, and the user-friendly design of the mobile application. The application will not require real-time connections, reducing the need for highly powerful servers. Instead, it will focus on efficient processing and accurate dubbing outputs.The required computational resources for AI processing can be managed within the mobile platforms and through cloud-based services for more intensive tasks.

Given the manageable technical requirements and the availability of necessary resources, the project is deemed feasible.

1.7 System Requirement

Users will be able to:

1. Register an account in the system.

2. Log in using their credentials.

3. Upload audio or video content for dubbing.

4. Record your voice.

5. Upload Recorded content for dubbing.

6. Receive dubbed audio output in Arabic.

7. Select alternative output voices, including celebrity voices.

8. View dubbed text output.

9. Edit the text output of the dubbed content for accuracy.

10. Create playlists and organize dubbed projects into customizable folders.

11. Download the dubbed audio to the local storage device

Admin will be able to:

1. Log in using their credentials.

2. View statistical studies and reports

3. View list of all users.

4. Add new user accounts.

5. Update existing user profiles.

6. Delete user accounts.

7. Access and review user history

8. Listen to both original and dubbed audio for human evaluation purposes.

9. View the text output of dubbed content

10. Log Out

Chapter 2: Theoretical Study

2.1 Deep Learning

Deep learning is the subset of machine learning methods which is based on artificial neural networks with representation learning. The adjective “deep” refers to the use of multiple layers in the network. Methods used can be either supervised, semi-supervised or Unsupervised. Deep-learning architectures such as deep neural networks, deep belief networks, recurrent neural networks, convolutional neural networks and transformers have been applied to fields including computer vision, speech recognition, natural language processing, machine translation, bioinformatics, drug design, medical image analysis, climate science, material inspection and board game programs, where they have produced results comparable to and in some cases surpassing human expert performance. Artificial neural networks (ANNs) were inspired by information processing and distributed communication nodes in biological systems. ANNs have various differences from biological brains. Specifically, artificial neural networks tend to be static and symbolic, while the biological brain of most living organisms is dynamic (plastic) and analog. ANNs are generally seen as low-quality models for brain function.[30]

2.1.1 DL Algorithms

2.1.1.1 Long Short-Term Memory:

(LSTM) network is a recurrent neural network (RNN), aimed to deal with the vanishing gradient problem present in traditional RNNs. Its relative insensitivity to gap length is its advantage over other RNNs, hidden Markov models and other sequence learning methods. It aims to provide a short-term memory for RNN that can last thousands of timestamps, thus “long short-term memory”. It is applicable to classification, processing and predicting data based on time series,

such as in handwriting, speech recognition, machine translation, speech activity detection, robot control, video games, and healthcare.[31]

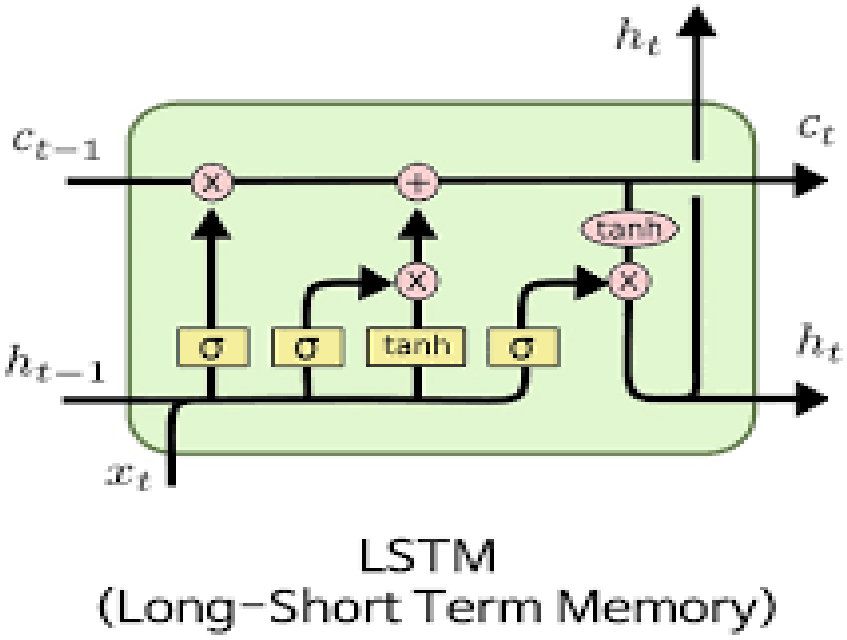


Figure [1]

2.1.1.2 Bidirectional Long Short-Term Memory (BiLSTM):

Bidirectional Long Short-Term Memory (BiLSTM) networks are an advanced form of LSTMs used in natural language processing (NLP). They enhance the standard LSTM model by processing sequences in both forward and backward directions, thus capturing information from both past and future contexts within the data. This dual processing allows BiLSTMs to better understand the structure and meaning of text, leading to improved performance in tasks such as text classification, sentiment analysis, and machine translation. [32]

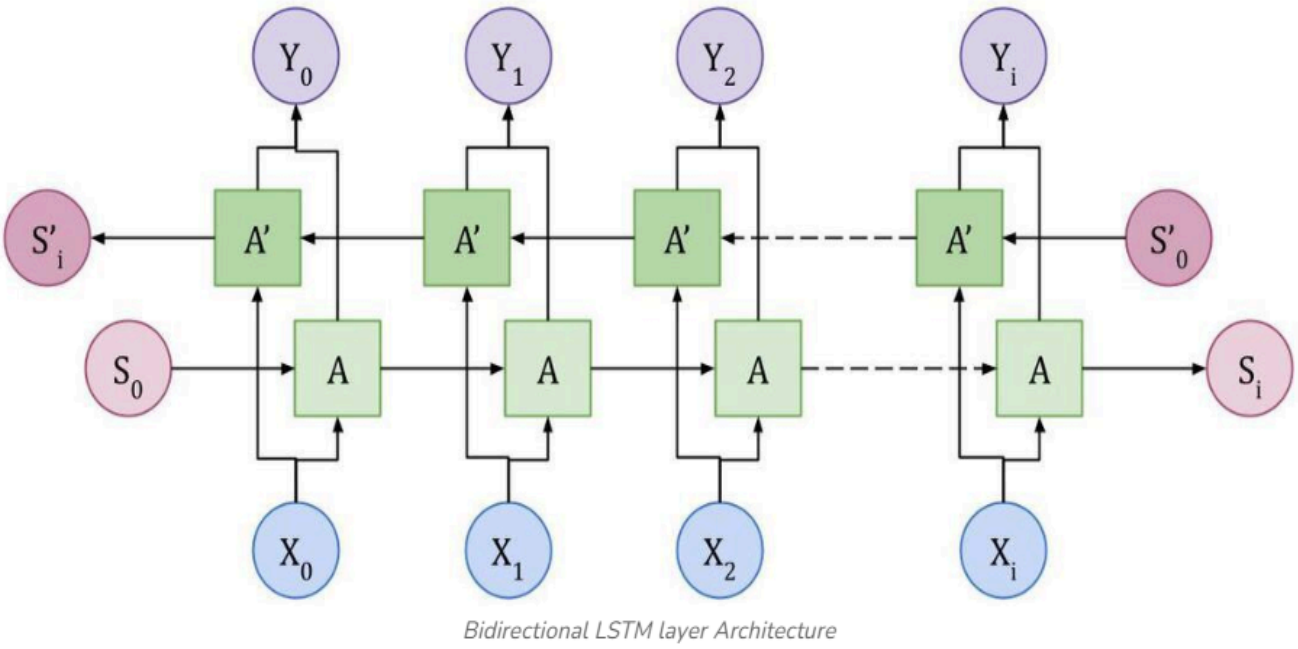


Figure [2]

2.1.1.3 Convolutional Neural Network (CNN):

CNN is a regularized type of feed-forward neural network that learns feature engineering by itself via filters (or kernel) optimization. Vanishing gradients and exploding gradients, seen during backpropagation in earlier neural networks, are prevented by using regularized weights over fewer connections. For example, for each neuron in the fully-connected layer 10,000 weights would be required for processing an image size 100x100 pixels. However, applying

cascaded convolution (or cross-correlation) kernels, only 25 neurons are required to process 5x5-sized tiles. Higher-layer features are extracted from wider context windows, compared to lower-layer features.[33]

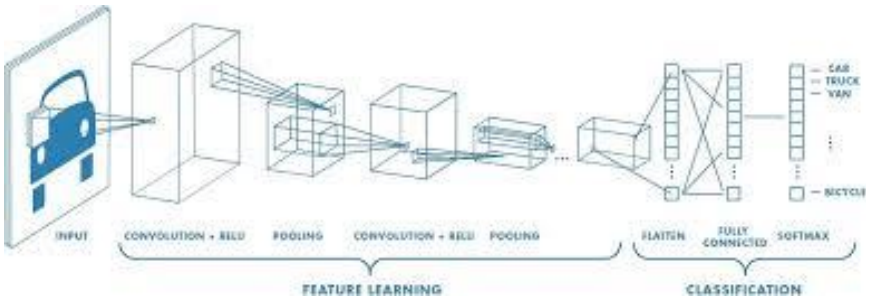


Figure [3]

2.2 Speech Translation

Speech translation is the process by which conversational spoken phrases are instantly translated and spoken aloud in a second language . This differs from phase translation , which is where the system only translates a fixed and finite set of phrases that have been manually entered into the system.Speech translation technology enables speakers of different languages to communicate. The first time speech translation was observed was during the 1983 ITU Telecom World (Telecom'83), when NEC Corporation demonstrated it as a proof of concept. In 1993, the ATR, Carnegie Melon University (CMU), and Siemens collaborated on a speech translation experiment that involved three locations around the world: the ATR, CMU, and Siemens.[9]

2.3 Speech Translation Technique

Speech-to-speech translation employs several advanced techniques to bridge the language divide effectively. One of the foundational methods is Rule-based Translation, which relies on predefined grammatical and vocabulary rules to convert text from one language to another. This was followed by Example-based Translation, leveraging pairs of sentences in source and target languages to generate translations through pattern matching. The field saw significant advancements with Statistical Machine Translation (SMT), pioneered by IBM, which uses statistical models derived from bilingual text corpora. A more refined approach, Phrase-based SMT, focuses on translating phrases rather than individual words, capturing more context and improving accuracy. Recently, Neural Machine Translation (NMT) has revolutionized the field, utilizing deep learning to handle whole sentences, thereby enhancing the capture of context and nuances. These translation techniques are supported by Automatic Speech Recognition (ASR), which converts spoken language into text, and Text-to-Speech (TTS) Synthesis, which transforms the translated text back into spoken language, ensuring a natural-sounding output. Additionally, integrated pipelines that combine ASR, MT, and TTS processes are crucial for achieving real-time speech-to-speech translation, thus making multilingual communication more accessible and practical for various applications. [10]

2.3.1 Voice Recognition (Speech-to-Text)

Voice recognition software on computers requires analog audio to be converted into digital signals, known as analog-to-digital (A/D) conversion. For a computer to decipher a signal, it must have a digital database of words or syllables as well as a quick process for comparing this

data to signals. The speech patterns are stored on the hard drive and loaded into memory when the program is run. A comparator checks these stored patterns against the output of the A/D converter -- an action called pattern recognition.

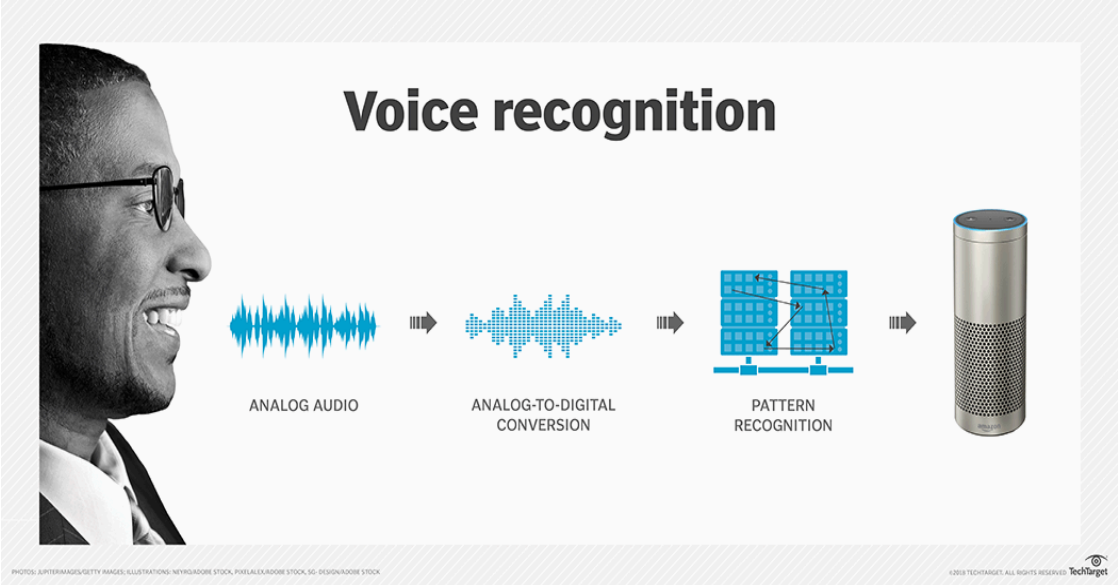


Figure [4]

Voice recognition systems analyze speech through one of two models: the hidden Markov model and neural networks. The hidden Markov model breaks down spoken words into their phonemes, while recurrent neural networks use the output from previous steps to influence the input to the current step. [11]

2.3.2 Machine Translation

In the age of advanced technology, Machine Translation has become an integral aspect of language learning and global communication. This article will provide an in-depth understanding of Machine Translation, beginning with a comprehensive introduction to the term, followed by an analysis of linguistic terms associated with Machine Translation. Furthermore, the development and historical background of this innovative technology will be discussed. As you delve deeper into the topic, you will also be introduced to various types and approaches of Machine Translation, including rule-based, statistical, and neural systems, as well as direct,

transfer, and interlingua methodologies. In the latter sections, the practical applications and limitations of Machine Translation will be explored alongside a comparison with computer-assisted translation and human translation. Stay tuned to equip yourself with valuable knowledge on this influential language technology.

Machine Translation (MT) is a subfield of computational linguistics that focuses on the automated translation of text or speech from one language to another. The primary goal of machine translation is to simplify and speed up the process of translating content while maintaining a high level of accuracy. MT systems can be classified into three main types: Rule-Based Machine Translation (RBMT), Statistical Machine Translation (SMT), and Neural Machine Translation (NMT). [12]

On 7 January 1954 the Georgetown–IBM experiment was held in New York at the head office of IBM. This was the first public demonstration of a machine translation system. The demonstration was widely reported in the newspapers and garnered public interest.

2.3.3 Speech Synthesis (Text-to-Speech)

Speech synthesis is the artificial production of human speech. A computer system used for this purpose is called a speech synthesizer and can be implemented in software or hardware products. A text-to-speech (TTS) system converts normal language text into speech; other systems render symbolic linguistic representations like phonetic transcriptions into speech.

Synthesized speech can be created by concatenating pieces of recorded speech that are stored in a database. Systems differ in the size of the stored speech units; a system that stores phones or diphones provides the largest output range, but may lack clarity. For specific usage domains, the

storage of entire words or sentences allows for high-quality output. Alternatively, a synthesizer can incorporate a model of the vocal tract and other human voice characteristics to create a completely "synthetic" voice output.

A text-to-speech system (or "engine") is composed of two parts:[3] a front-end and a back-end. The front-end has two major tasks. First, it converts raw text containing symbols like numbers and abbreviations into the equivalent of written-out words. This process is often called text normalization, pre-processing, or tokenization. The front-end then assigns phonetic transcriptions to each word, and divides and marks the text into prosodic units, like phrases, clauses, and sentences. The process of assigning phonetic transcriptions to words is called text-to-phoneme or grapheme-to-phoneme conversion. Phonetic transcriptions and prosody information together make up the symbolic linguistic representation that is output by the front-end. The back-end—often referred to as the synthesizer—then converts the symbolic linguistic representation into sound. In certain systems, this part includes the computation of the target prosody (pitch contour, phoneme durations). Speech synthesis systems use two basic approaches to determine the pronunciation of a word based on its spelling, a process which is often called text-to-phoneme or grapheme-to-phoneme conversion (phoneme is the term used by linguists to describe distinctive sounds in a language).

The first computer-based speech-synthesis systems originated in the late 1950s. Noriko Umeda et al. developed the first general English text-to-speech system in 1968, at the Electrotechnical Laboratory in Japan.[13]

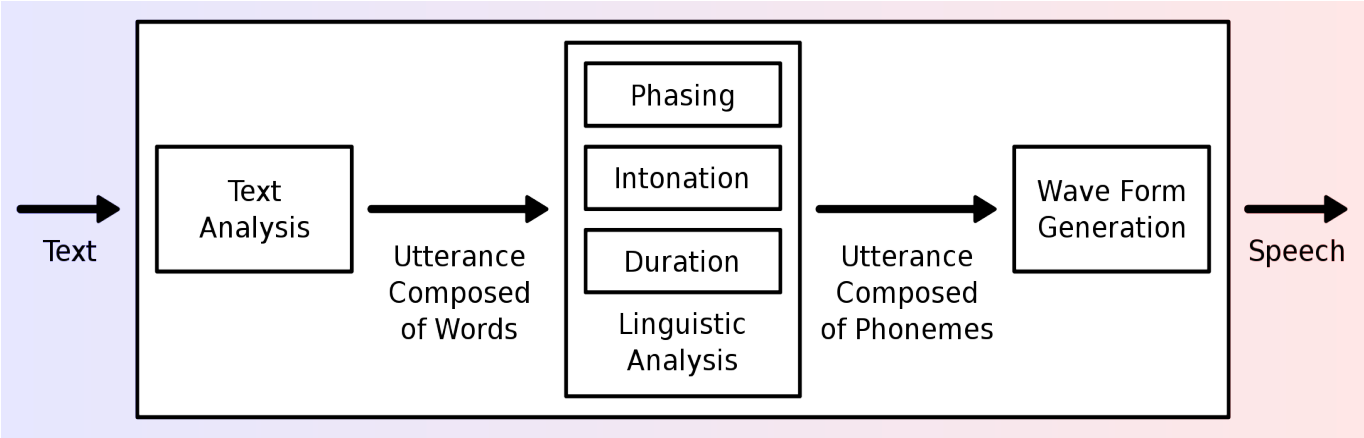


Figure [5]

2.3.4 Voice Cloning

Voice cloning is the creation of an artificial simulation of a person's voice.Today's AI software methods are capable of generating synthetic speech that closely resembles a targeted human voice. In some cases, the difference between the real and fake voice is imperceptible to the average person. [14]

In fact, AI voice cloning was invented all the way back in 1998! Since then, the technology has continued to develop and improve, and we're now at a point where many AI voice clips you hear are indistinguishable from the real thing. [15]

Voice cloning is usually created using a text to speech system that converts written text into human speech. This is usually done using artificial intelligence and machine learning algorithms. Like other types of generative AI , the process of creating voice cloning systems

begins with collecting huge amounts of data to create a dataset of the individual's voice samples. These voice samples or recordings should contain a variety of voices, accents, tones, and expressions to represent different voices, nuances, and situations. These samples are then organized , labeled , and fed to the AI models where they will be used. This is the first stage of creating a working voice cloning model. [16]

Chapter 3: Literature Review

3.1 English Speech-to-Text

3.1.1 Datasets

The datasets mentioned provide a broad spectrum of resources for speech recognition and linguistic research. Common Voice, with over 9,000 hours of speech in 70+ languages, is a crowdsourced project by Mozilla aiming to make speech recognition accessible globally. TIMIT, containing 5.4 hours of American English from 630 speakers, is essential for phonetic and acoustic research. LibriSpeech offers around 1,000 hours of English speech from audiobooks, ideal for training ASR systems. AISHELL, available in two versions with 178 and 1,000 hours, focuses on Mandarin Chinese. Switchboard Hub500 and the Fisher Corpus, with 300 and 2,000 hours respectively, provide extensive conversational English data for various linguistic tasks. Baidu's datasets, primarily in Mandarin, contribute to speech recognition advancements in Chinese. The WSJ corpus, with 80 hours of read English news text, serves as a benchmark for continuous speech recognition. Lastly, the SPGISpeech Corpus, with 5,000 hours of English speech, focuses on financial and business content, enhancing domain-specific speech recognition capabilities. These datasets collectively cater to diverse linguistic and application-specific needs in the field of speech technology.

3.1.1 State of Art

Ravanelli, M., Parcollet, T., Plantinga, P., Rouhe, A., Cornell, S., Lugosch, L., Subakan, C., Dawalatabad, N., Heba, A., Zhong, J., Chou, J., Yeh, S., Fu, S., Liao, C., Rastorgueva, E., Grondin, F., Aris, W., Na, H., Gao, Y., De Mori, R., and Bengio, Y. (2021). SpeechBrain: A

comprehensive toolkit for speech-to-text applications. In this research, the authors tackle the challenges of speech recognition by leveraging advanced deep learning models and architectures. Utilizing state-of-the-art techniques such as CRDNN, Transformer, and wav2vec 2.0, SpeechBrain achieves impressive performance across several benchmarks. Notably, the transformer-based models reach a word error rate (WER) of 2.46% on the LibriSpeech test-clean subset, while models employing a combination of CTC and attention with a wav2vec 2.0 encoder achieve a phone error rate (PER) of 8.04% on the TIMIT dataset. This study significantly contributes to the field by providing a flexible, high-performance toolkit that simplifies the development and deployment of speech processing pipelines. The open-source nature of SpeechBrain, released under the Apache 2.0 license, ensures that it can be widely adopted and contribute to further advancements in speech recognition research and applications.

Hannun, A., Case, C., Casper, J., Catanzaro, B., Diamos, G., Elsen, E., Prenger, R., Satheesh, S., Sengupta, S., Coates, A., & Ng, A. Y. (2014) address the challenge of end-to-end speech recognition, presenting a system named "Deep Speech" that utilizes deep learning to simplify and improve traditional speech systems. The authors employ a recurrent neural network (RNN) architecture optimized for multiple GPUs and extensive data synthesis techniques, eliminating the need for phoneme dictionaries or hand-designed components for background noise and speaker variation. The methodology involves preprocessing speech signals into spectrograms and training the RNN to generate character probabilities for transcription. The study, based on a variety of datasets such as: Fisher corpus, Baidu , WSJ & Switchboard Hub5’00 corpus that

achieves a state-of-the-art error rate of 16.0%. The research contributes to the field of speech recognition by demonstrating the efficacy of end-to-end deep learning models in handling challenging noisy environments and outperforming traditional systems.

O'Neill, P. K., Lavrukhin, V., Majumdar, S., Noroozi, V., Zhang, Y., Kuchaiev, O., Balam, J., Dovzhenko, Y., Freyberg, K., Shulman, M. D., Ginsburg, B., Watanabe, S., and Kucsko, G. (2021) address the challenge of fully formatted end-to-end speech recognition by presenting a novel task where acoustic models generate text with comprehensive orthographic features directly from audio. The authors introduce SPGISpeech, a corpus consisting of 5,000 hours of transcribed financial audio from earnings calls. This corpus is distinguished by its professionally transcribed, fully formatted text and a wide range of speech conditions. The study demonstrates the feasibility of this approach using Conformer-based models combined with the SpecAugment feature, achieving a character error rate (CER) of 1.0% and a word error rate (WER) of 2.3%. This research significantly contributes to the field of speech-to-text (STT) by simplifying the transcription process and enhancing accuracy through rich acoustic information. Moreover, the SPGISpeech corpus is provided for free academic use, thereby expanding the resources available for STT model development.

Miao, Y., Gowayyed, M., & Metze, F. (2015). Eesen: End-to-end speech recognition using deep RNN models and WFST-based decoding. This research introduces the Eesen framework, which simplifies the ASR pipeline using deep bidirectional recurrent neural networks (RNNs) with Long Short-Term Memory (LSTM) units for acoustic modeling. By employing the connectionist temporal classification (CTC) objective function, Eesen removes the need for pre-generated

frame labels. A distinctive feature of Eesen is its WFST-based decoding approach, which integrates lexicons and language models efficiently.Using the Wall Street Journal (WSJ) corpus, Eesen achieves a word error rate (WER) of 7.34% with an expanded vocabulary and re-trained trigram language model, demonstrating superior performance compared to other end-to-end systems. This contribution significantly advances the field of speech recognition by providing an efficient and high-performing ASR framework that is open-source, allowing for continuous improvement and widespread adoption.

Paper Title	Features	Methods	Datasets	Accuracy	Year
SpeechBrain: A General-Purpose Speech Toolkit [17]	80 fbanks 40 fbanks	CRDNN(CNN LSTM DNN) TransformerASR ECAPA-TDNN X-vector GRU Beam search LM ContextNet	-Common Voice - TIMIT -LibriSpeech -AISHELL	**2.46** 9.86 13.34 (WER)	2021
Deep Speech: Scaling up end-to-end speech recognition [18]	Spectrogram Energy Term	RNN Bidirectional recurrent layer	Switchboard Hub500 Fisher corpus Baidu WSJ	**16.0% (WER)**	2014
SPGISpeech: 5,000 hours of transcribed financial audio for fully formatted end-to-end speech recognition [19]	SpecAugment	Conformer	SPGISpeech Corpus	**2.3% (wer)** **1.0% (cer)**	2021
EESEN: END-TO-END SPEECH RECOGNITION USING DEEP RNN MODELS AND WFST-BASED DECODING [20]	40-dimensional filterbank features	bi-directional LSTM WFST DNN	WSJ corpus	**7.34% (WER)**	2015

End-to-End Mandarin Speech Recognition Combining CNN and BiLSTM	MFCC	CNN BiLSTM	AISHELL-1	**19.2% (WER)**	2019

Papers of Speech to text

3.2 Arabic Text-to-Speech

3.2.1 Datasets:

The datasets mentioned are crucial for advancing research in Arabic and general speech synthesis, each tailored to specific and broad applications. The "Half-Diacritized Lexicon of Sample Arabic Words" comprises an allophone/diphone database with pre-recorded speech units essential for synthesizing speech from Arabic text inputs. Although this database supports the production of intelligible and natural-sounding speech, it requires further enhancements to improve its quality and naturalness. The "Modern Standard Arabic Phonetically-Balanced Speech Dataset" includes 3,000 sentences, recorded in a professional studio to ensure a phonetically balanced and diverse range of expressions in Modern Standard Arabic, optimizing the phonetic context within scripts. The "Multi-Genre Broadcast (MGB2)" dataset enriches models' capability to handle various dialects and informal styles of Arabic with its extensive collection of Arabic broadcasts. Additionally, the "Arabic Speech Corpus (ASC)" and the "Classical Arabic Text-to-Speech Corpus (ClArTTS)" provide structured resources for developing TTS systems in standard and classical Arabic, respectively. Expanding beyond Arabic, the "AISHELL-3" dataset, which includes approximately 85 hours of Mandarin Chinese speech, is aimed at training more inclusive speech recognition and synthesis systems. The "LibriTTS" dataset, derived from the renowned "LibriSpeech" dataset, features 585 hours of clean, segmented speech from English audiobooks, specifically designed for TTS applications. Together, these datasets facilitate significant advancements in language modeling and speech synthesis across multiple languages and dialects.

3.2.2 State of Art:

Hamad, M., & Hussain, M. (2011) address the development of an Arabic Text-To-Speech (TTS) synthesizer, highlighting the need for advanced speech synthesis technologies in languages less commonly researched like Arabic. The authors introduce a system that employs allophone/diphone concatenation methods, optimized for processing complex linguistic inputs such as abbreviations, numbers, and special characters. This system is designed to output speech from text using a synthesized male voice. The architecture allows for adjustments to accommodate other verbal languages with minor modifications to its phonetic database, aimed at broadening its applicability. The methodology involves a dual-module approach: a text/linguistic analyzer and a synthesizer core, ensuring precise pronunciation and natural intonation. The study tests this system with various types of Arabic text, aiming to establish a foundational system that can be expanded with additional voices and linguistic nuances. The research advances the field of TTS by demonstrating the system's potential to support educational and accessibility applications for Arabic speakers, and its adaptability to other languages, thereby enhancing the inclusivity and reach of TTS technology.

Youssef, A., & Emam, O. (2004) present an advanced Arabic Text-to-Speech (TTS) system developed at IBM Egypt's Human Language Technologies laboratory. This state-of-the-art system is built on IBM's trainable unit-selection based concatenative speech synthesizer. The paper provides a comprehensive review of the system components with a focus on those features that specifically cater to the Arabic language. The TTS system capitalizes on a large speech database to generate high-quality,

natural-sounding speech, addressing the gap in Arabic speech synthesis technologies compared to other languages like English. The system architecture includes modules for text normalization, phonological analysis, and prosodic planning, leading to synthesized speech that demonstrates significant improvements in naturalness and intelligibility. The authors conclude with results from subjective tests which show promising mean opinion scores for the synthesized speech, affirming the system's effectiveness. Future work aims to further refine the system's capabilities and expand its linguistic adaptability.

Toyin, H. O., Djanibekov, A., Kulkarni, A., & Aldarmaki, H. (2023) introduce ArTST, a groundbreaking pre-trained Arabic text and speech transformer, adapted from the SpeechT5 framework. This model is specifically designed for Modern Standard Arabic (MSA) and focuses on enhancing open-source speech technologies for Arabic. It is pre-trained from scratch on substantial MSA speech and text data, and fine-tuned for various tasks including Automatic Speech Recognition (ASR), Text-To-Speech synthesis (TTS), and spoken dialect identification. The experimental results demonstrate that ArTST matches or surpasses the current state-of-the-art in these fields, showcasing its effective generalization, particularly in low-resource TTS applications. This model also offers promising directions for future enhancements, including potential adaptations for dialectal and code-switched Arabic. The researchers have made the pre-trained and fine-tuned models publicly available for research purposes, significantly contributing to advancements in Arabic speech technology.

Zhou, Y., Song, C., Li, X., Zhang, L., Wu, Z., Bian, Y., Su, D., & Meng, H. (2022) introduce a novel approach to zero-shot speaker adaptation in text-to-speech synthesis by incorporating content-dependent fine-grained speaker embedding. This method aims to capture not only the global speaker characteristics but also the subtle pronunciation nuances related to phoneme content, which are often overlooked in fixed-length speaker embeddings. By using a reference attention module that models content relevance, the system dynamically adjusts the fine-grained speaker embedding for each phoneme in the input text, improving the speaker similarity in synthesized speech, particularly for unseen speakers. This technique represents a significant advance in generating more personalized and natural-sounding synthetic speech by effectively transferring individual pronunciation styles and habits from a reference speech to synthesized outputs. The findings suggest enhanced speaker similarity and more authentic voice cloning in TTS systems, showing promising results for practical applications where diverse speaker voices need to be accurately mimicked without extensive adaptation data.

Lee, J. Y., Jeong, M., Kim, M., Lee, J.-H., Cho, H.-Y., & Kim, N. S. (2024) introduce an innovative two-stage text-to-speech (TTS) framework designed for high-fidelity speech synthesis, leveraging two types of discrete tokens—semantic and acoustic. The framework consists of an Interpreting module that converts text into semantic tokens, capturing linguistic content and alignment, and a Speaking module that translates these tokens into acoustic tokens to produce the final speech output, embodying the desired timbre and acoustic properties. The Interpreting module utilizes a Token Transducer for robust alignment, while the Speaking module incorporates a Group Masked Language Model (G-MLM) using a Conformer-based architecture for enhanced efficiency. This method achieves superior performance in

zero-shot TTS scenarios, particularly excelling in speech quality and speaker similarity, making it a significant advancement in the realm of synthetic speech technologies.

Paper Title	Features	Methods	Datasets	Accuracy	Year
Arabic Text-To-Speech Synthesizer [21]	-Text/linguistic analyzer -synthesizer core	Allophone/Diphone Concatenation	half-diacritized lexicon of sample Arabic words	Intelligibility: Acceptable, Naturalness: Needs improvement, Sound quality: Acceptable, Pronunciation: Consonant-consonant junctions not satisfactory (MOS)	2011
An Arabic TTS System Based on the IBM Trainable Speech Synthesizer [22]	MFCC Pitch marks Pitch Synchronous (PS) Frames	HMM	Modern Standard Arabic Phonetically -Balanced Speech Dataset	Intelligibility: 4.0 Naturalness: 3.4, voice quality: 3.65 (MOS)	2004
ArTST: Arabic Text and Speech Transformer [23]	-Log Mel Filterbanks -MFCCs -80-dimensional Log Mel Filterbanks	SpeechT5 Transformer HiFi-GAN vocoder	Multi-Genre Broadcast (MGB2) Arabic Speech Corpus (ASC) Classical Arabic Text-to-Speech Corpus (ClArTTS)x	12.51% (WER) 3.60% (CER)	2023
Content-Dependent Fine-Grained Speaker Embedding for Zero-Shot Speaker Adaptation in Text-to-Speech Synthesis [24]	Mel-spectrogram	CDFSE GSE CLS Attentron HIFI-GUN	AISHELL-3	MOS: 3.59 (seen speakers), 3.54 (unseen speakers)	2022

High Fidelity Text-to-Speech Via Discrete Tokens Using Token Transducer and Group Masked Language Model [25]	-Semantic and acoustic tokens -Mel Spectrogram	-G-MLM -Group-Iterative Parallel Decoding (G-IPD) -Cross Attention-based Generator -Token Transduce	LibriTTS	MOS: 3.94	2024

Papers of Text to speech

3.3 Arabic Text-to-Speech & Voice Cloning

3.3.1 DataSets

The Voice Bank Corpus. University of Edinburgh, Centre for Speech Technology Research (CSTR). The CSTR VCTK Corpus, also known as the Voice Bank Corpus, comprises speech data from 109 native speakers of English with various accents. Recorded at 48 kHz, the dataset includes both male and female voices along with text transcriptions. This corpus is intended for research and development in speech synthesis and speaker recognition, offering a diverse set of voices and accents to enable robust model training and evaluation in the field of speech technology.

LibriSpeech: An ASR Corpus Based on Public Domain Audio Books. In 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 5206-5210). IEEE. The LibriSpeech dataset is a comprehensive collection of approximately 1,000 hours of English read speech derived from audiobooks that are part of the LibriVox project. Sampled at 16 kHz, the dataset includes text transcriptions and is designed to support the development and evaluation of automatic speech recognition (ASR) systems. LibriSpeech provides a rich resource for training and testing models in the speech processing community, facilitating advancements in ASR technologies.

Common Voice : A Massively-Multilingual Speech Corpus. Mozilla Common Voice. Common Voice is a large-scale, multilingual speech corpus developed by the Mozilla Foundation. This dataset includes voice recordings contributed by volunteers and spans multiple languages. It aims to support the development of open-source speech recognition technologies by providing diverse

voice data suitable for training and evaluating speech processing models. The Common Voice initiative promotes inclusivity and diversity in speech technology research, facilitating advancements in multilingual and accent-inclusive speech recognition systems.

3.3.2 State of Art

Coqui(7 June 2024) XTTS: a Massively Multilingual Zero-Shot Text-to-Speech Model :This research on XTTS leverages advanced neural network architectures to deliver state-of-the-art performance in zero-shot, multi-speaker TTS applications. Utilizing an autoencoder(VQ-VAE), encoder (GPT-2), and decoder (based on the HiFi-GAN vocoder), the model generates text-to-speech with voice cloning capabilities.

The evaluation compared the XTTS model with several state-of-the-art (SOTA) zero-shot text-to-speech (ZS-TTS) models, including StyleTTS 2, Tortoise, YourTTS, HierSpeech++

and Mega-TTS 2. The comparison used 240 sentences from the FLORES+ dataset and 20 speakers from the DAPS dataset. Evaluation metrics included Naturalness Mean Opinion Score (nMOS) using UTMOS, Speaker Encoder Cosine Similarity (SECS) for voice similarity, and Character Error Rate (CER) for pronunciation accuracy. Subjective user preference scores were also collected. All evaluation code and audio samples are available in the ZS-TTS-Evaluation repository for reproducibility.

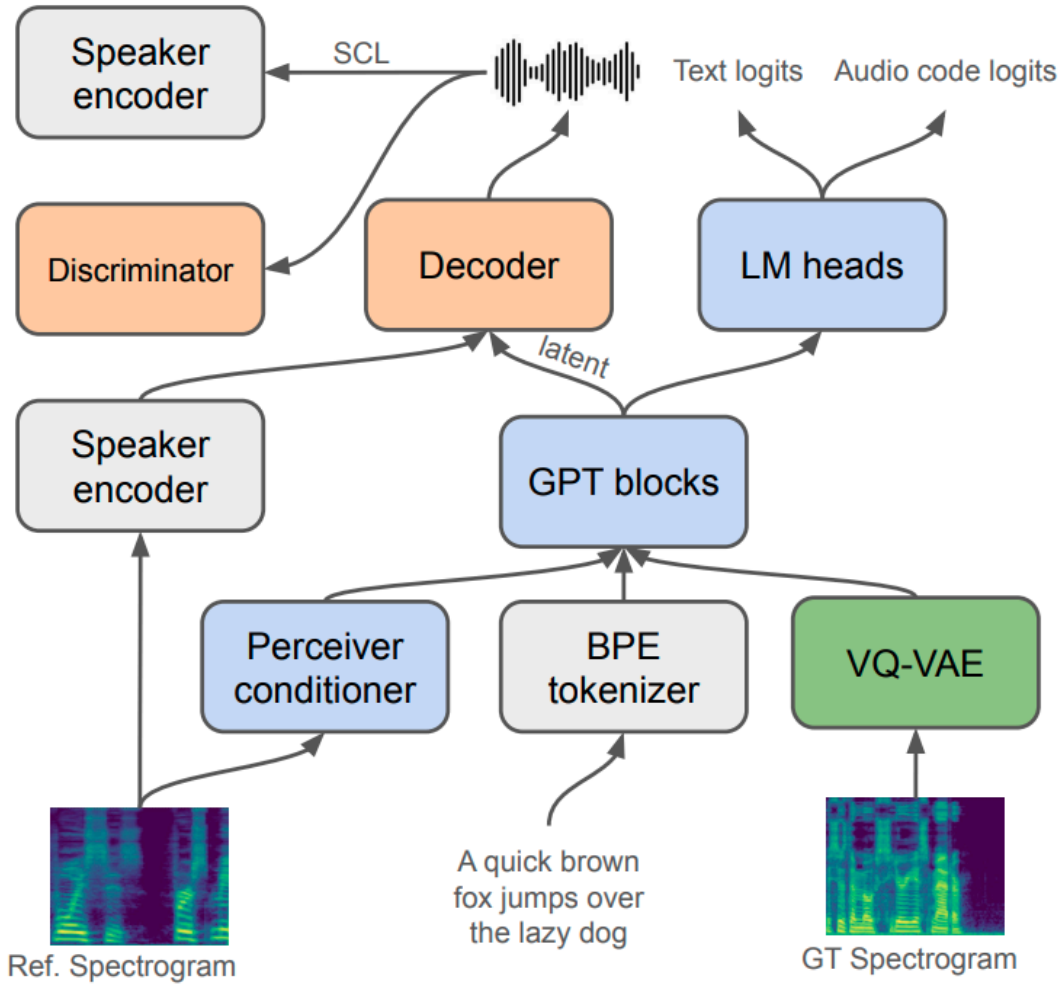

Speaker encoder
SCL
Text logits
Audio code logits
Discriminator
Decoder
LM heads
latent
Speaker encoder
GPT blocks
Perceiver conditioner
BPE tokenizer
VQ-VAE
Ref. Spectrogram
A quick brown fox jumps over the lazy dog
GT Spectrogram

Figure [6]

“**VQ-VAE:** A Vector Quantised-Variational AutoEncoder (VQ-VAE) with 13M parameters receives a mel-spectrogram as input and encodes each frame with 1 codebook consisting of 8192 codes at a 21.53 Hz frame rate. The architecture and training procedure of VQ-VAE is the same as the one used in [1]; however, after VQ-VAE training we have filtered the codebook keeping only the first 1024 most frequent codes. In preliminary experiments, we verified that filtering the less frequent codes improved the model’s expressiveness.

Encoder: The GPT-2 encoder is a decoder-only transformer that is composed of 443M parameters, similar to [1]. It receives as inputs text tokens obtained via a 6681-token custom Byte-Pair Encoding (BPE) [2] tokenizer and as output predicts the VQ-VAE audio codes. The GPT-2 encoder is also conditioned by a Conditioning Encoder, described below, that receives mel-spectrograms as input and produces 32 1024-dim embeddings for each audio sample. The Conditioning Encoder is composed of six 16-head Scaled Dot-Product Attention layers followed by a Perceiver Resampler [3] to produce a fixed number of embeddings independently of the input audio length. Note that in [1] the authors didn’t use the Perceiver Resampler, instead, they used only a single 1024-dim embedding to condition the GPT-2 encoder. In our preliminary experiments, we noticed that in massive multilingual training, the use of a single embedding leads to a decrease in the model’s speaker cloning capability. We also have romanized the texts before tokenization for the Korean, Japanese, and Chinese languages using hangul-romanize4 , Cutlet5 , and Pypinyin6 packages respectively.

Decoder: The decoder is based on the HiFi-GAN vocoder [5] with 26M parameters. It receives the latent vectors out of the GPT-2 encoder. Due to the high compression rate of the VQ-VAE,

reconstructing the audio directly from the VQ-VAE codes leads to pronunciation issues and artifacts. To avoid this issue, we follow [1] and we have used the GPT-2 encoder latent space as input to the decoder instead of VQ-VAE codes. Our proposed decoder is also conditioned with speaker embedding from the H/ASP model [6]. The speaker embedding was added in each upsampling layer via linear projection. Inspired by [4], to improve the speaker similarity, we also added the Speaker Consistency Loss (SCL). To speed up inference we have trained the VQ-VAE and the encoder using 22.5 kHz audio signals. However, we train the decoder by upsampling the input vectors linearly to the correct length to produce 24khz audio"[0]

OpenVoice: Versatile Instant Voice Cloning. The paper presents a novel framework for instant voice cloning that is both versatile and efficient. Leveraging advanced deep learning techniques, the OpenVoice system can accurately replicate a speaker's voice with minimal data input, producing high-fidelity synthetic speech. The authors utilize prominent datasets, including LibriSpeech, VCTK, and CommonVoice, to train and evaluate the model, ensuring robustness across various accents and languages. The proposed method integrates a multi-speaker encoder, a fine-tuned vocoder, and an attention-based sequence-to-sequence model to achieve high-quality voice cloning. Key features of OpenVoice include its ability to generalize to unseen speakers, handle diverse acoustic conditions, and deliver low-latency performance. The method demonstrates significant improvements in voice cloning quality and speed, paving the way for enhanced applications in personalized speech synthesis, virtual assistants, and accessibility technologies.

Paper Title	Features	Methods	Datasets	Accuracy	Year
XTTS: a Massively Multilingual Zero-Shot Text-to-Speech Model [26]	Multilingual TTS Zero-Shot Voice Cloning Emotion recognition	Speaker Embedding Mel-spectrogram YourTTS & Tortoise Pre-trained models	Common Voice , LibriLight , LibriTTS-R	0.5425 (CER) 0.5007 (SECS)	2024
OpenVoice: Versatile Instant Voice Cloning	Voice Cloning TTS	HiFi-GAN VITS LSTM CNN	Common Voice , VCTK, LibriSpeech	4.57 (MOS)	2024

Papers of Text to speech & voice cloning

3.4 Similar Applications

Features: / Apps:	Speechify	BlipCut	AI Voice Dubbing (Our App)
Video Dubbing	✓	✓	✗
Audio content as text	✓	✓	✓
Selecting a suitable voice	✓	✓	✓
Organize as playlists	✗	✗	✓
Upload Audio/Video or Youtube link	✓	✓	✓
Multiple domains	✓	✓	✓
Multiple Languages	✓	✓	✗
Voice Recorder	✗	✗	✓
Free	✗	✗	✓

Chapter 4: System Analysis

4.1 ACID Theorem

The ACID properties have been highlighted to emphasize the reliability of transaction processing

in the system.

the following four properties are essential to ensure database transactions

are processed reliably:

- **Atomicity**: Each transaction is treated as a single unit, either fully succeeding or failing..
- **Consistency**: A transaction moves the database from one valid state to another,
- maintaining all rules.
- **Isolation**: Transactions are processed independently without interference from others.
- **Durability:** Once committed, transactions remain permanent, even after a system crash

The system's database uses a relational model (MySQL), which supports ACID properties ensuring reliable transaction processing.

4.2 Architectural Pattern

Architectural patterns are how to represent and use knowledge.

It is a specific solution for a specific set of concerns in software development.

These patterns are typically used within specific parts of a system to address design challenges.

4.3 Model View Control (MVC)

Our project employs the MVC (Model-View-Controller) architectural pattern, one of the most widely recognized and utilized patterns in system design. The primary goal of this pattern is to separate the data presentation from the actual data processing.

- **Model:**
    - This component is responsible for managing system data and associated processes.
    - Data is exposed to the client side via RESTful APIs.
- **View:**
    - This component handles displaying data to the user.
    - The Flutter framework is used for developing the client-side application.
    - The Flutter app consumes APIs provided by the Laravel backend to display data.
- **Control:**
    - This component manages user interactions and mediates between the View and the Model.
    - Laravel controllers process incoming requests, execute appropriate logic, and return responses.

User inputs from the Flutter app are sent to the Laravel backend via API calls, where they are processed, and appropriate data is returned.

4.4 Business Logic Component (BLoC)

BLoC is an architectural design pattern specifically implemented for Flutter applications.
It mainly helps to manage the separation of content in the application by dividing it into distinct layers: UI – BLoC – Data (received via APIs).

The main reason behind the usage of BLoC is to be able to set shared components & data between pages and widgets in one place and listen to their changes, which improves the writability & readability of the code and the overall performance.

BLoC Concept Example:

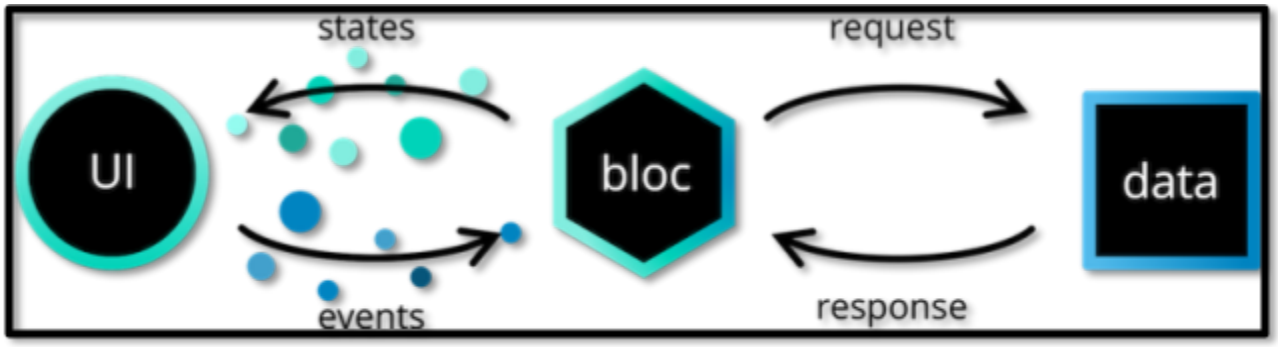


Figure [7]

4.5 Client-Server Architecture Style

In client server style, the system is organized and distributed into services, and each service provides a different job.

Clients use these services to achieve a specified objective.

In this architectural style, the client initiates the connection with the back-end side that will respond with data that satisfies the user's requests.

4.6 Functional Requirements

A User will be able to do the following:

1. Register & Login.

2. Edit personal information.
3. Upload media (audio – video) by choosing a file.
4. Upload Youtube link.
5. Record Voice.
6. Choose a dubbing voice.
7. Check previous dubbings.
8. Play dubbing.
9. Name dubbing.
10. View original or dubbed text.
11. Edit original or dubbed text.
12. Create a Playlist.
13. Organize Playlists (add dubbed audio files to preferred playlist).
14. Delete Playlists.
15. Delete account.
16. Logout.

<u>The Administrator will be able to do the following:</u>

1. Log in using their credentials.

2. View statistical information and reports
3. View list of all users.
4. Add new user accounts.
5. Update existing user profiles.
6. Delete user accounts.
7. View user history.
8. Listen to both original and dubbed audio for human evaluation purposes.
9. View the text output of original and dubbed content.
10. Logout.

4.7 Non-Functional Requirements

The Non-functional requirements define the system properties and its constraints.

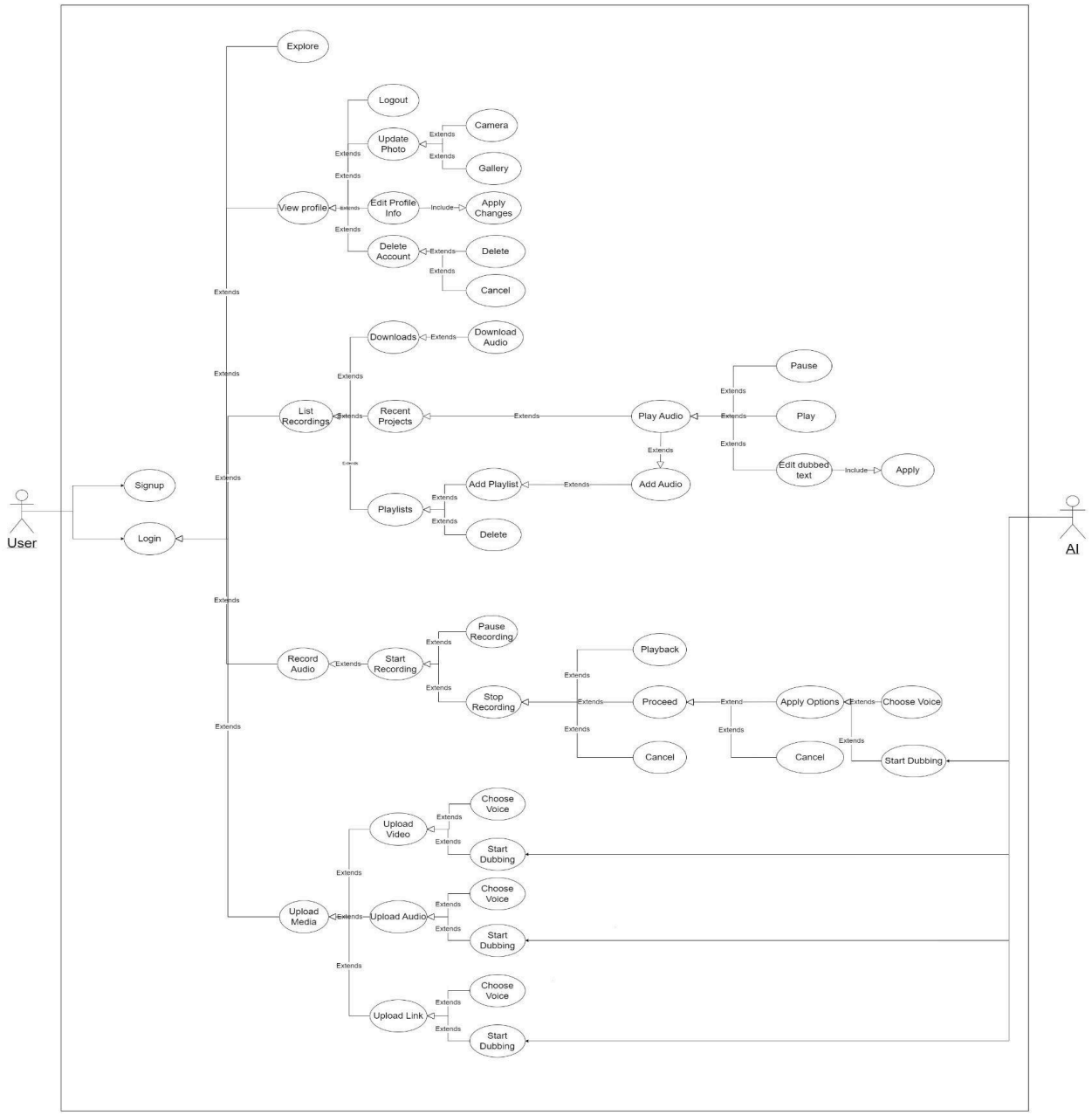
The importance of non-functional requirements can be the same as the functional requirements since it mainly monitors and assures the performance and the characteristics of the system to be.

- **Scalability:** The system handles an increasing number of users without compromising performance.
- **Reliability**: It specifies how likely the system or its element would run without a failure for a given period of time under predefined conditions.
- **Availability**: describes how likely the system is accessible to a user at a given point in time.
- **Security**:

- Authentication and Authorization: The system implements OAuth 2.0 Token, a token-based authentication mechanism for user access, providing secure generation, transmission, and storage of tokens and including proper verification methods.
- Password Hashing: User passwords are securely hashed using strong and standard hashing algorithms, ensuring the confidentiality and integrity of stored password data.
- Password Policies: The system follows a password policy, requiring users to create passwords that meet specific criteria (e.g., length, character types) during registration or password updates.

4.8 Use Case Diagram

4.8.1 User Use Case Figure [8]:



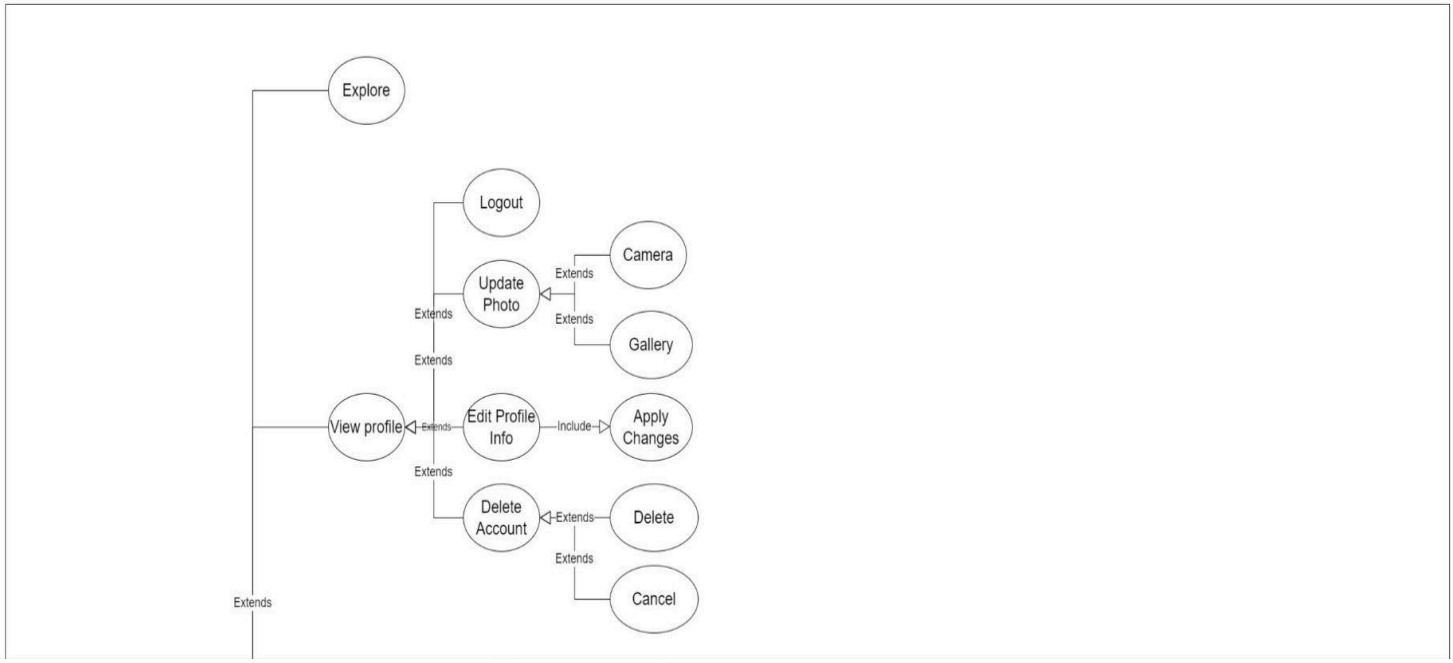


Figure [9] (Same Use Case Divided into three clear parts)

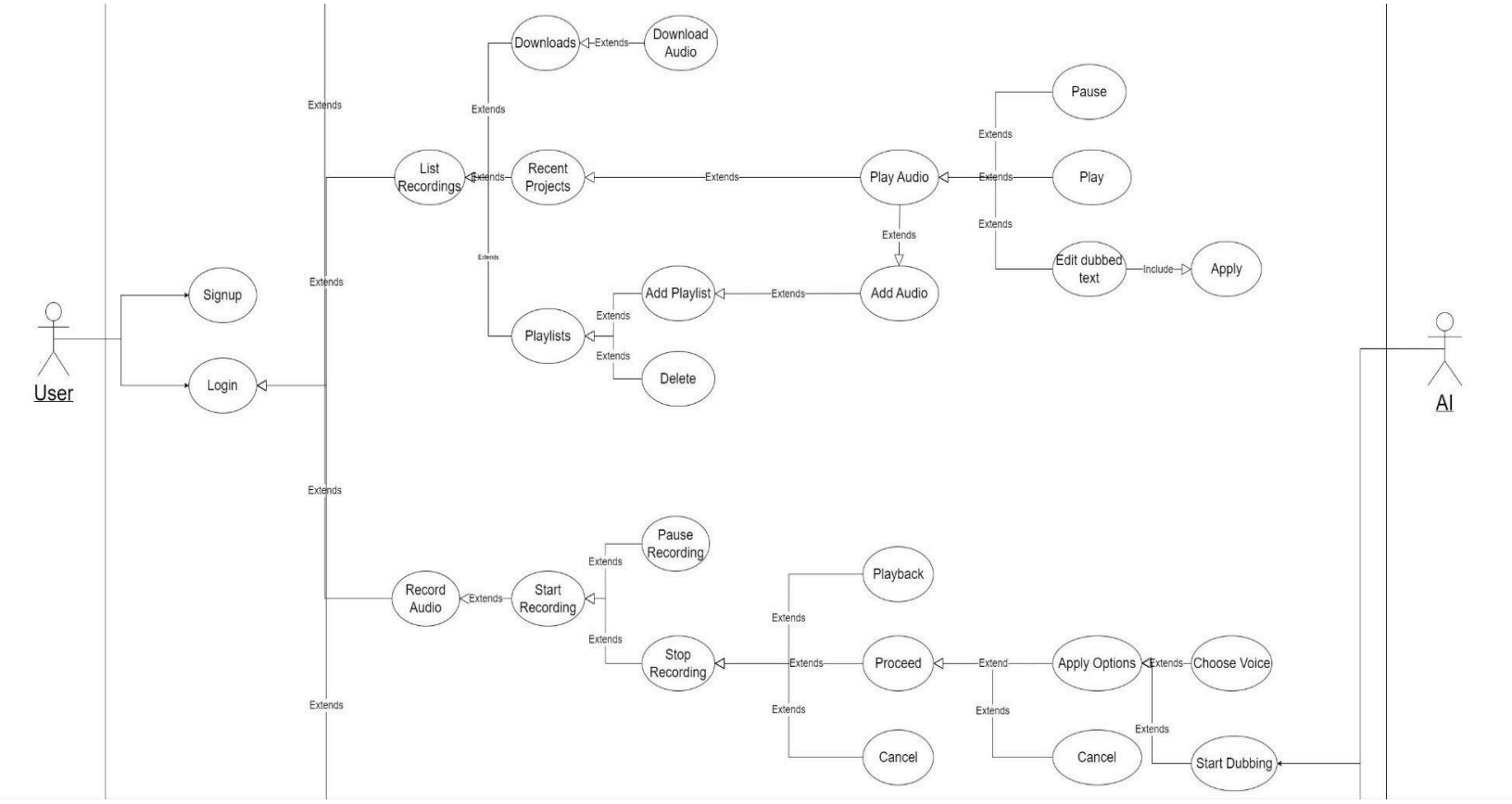


Figure [10] (Same Use Case Divided into three clear parts)

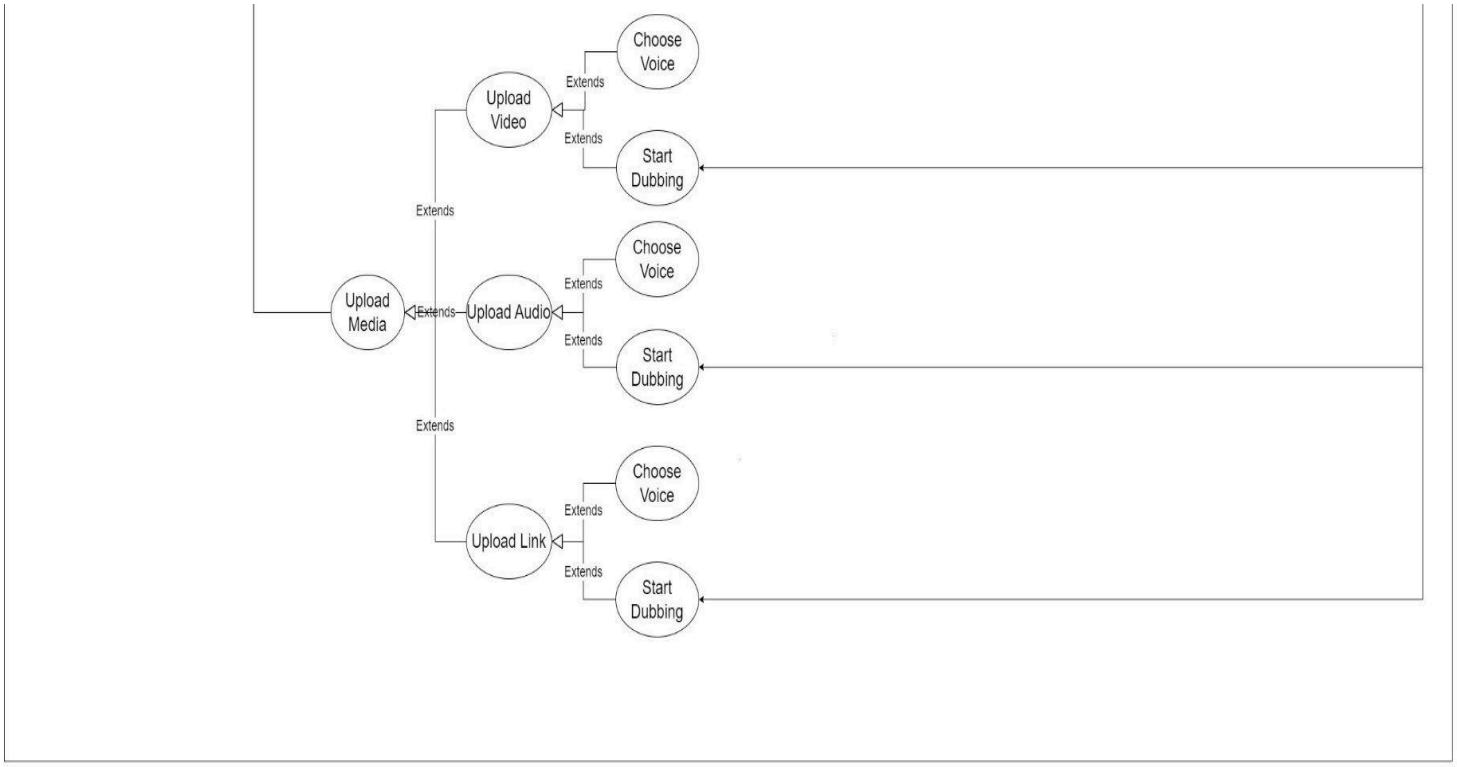
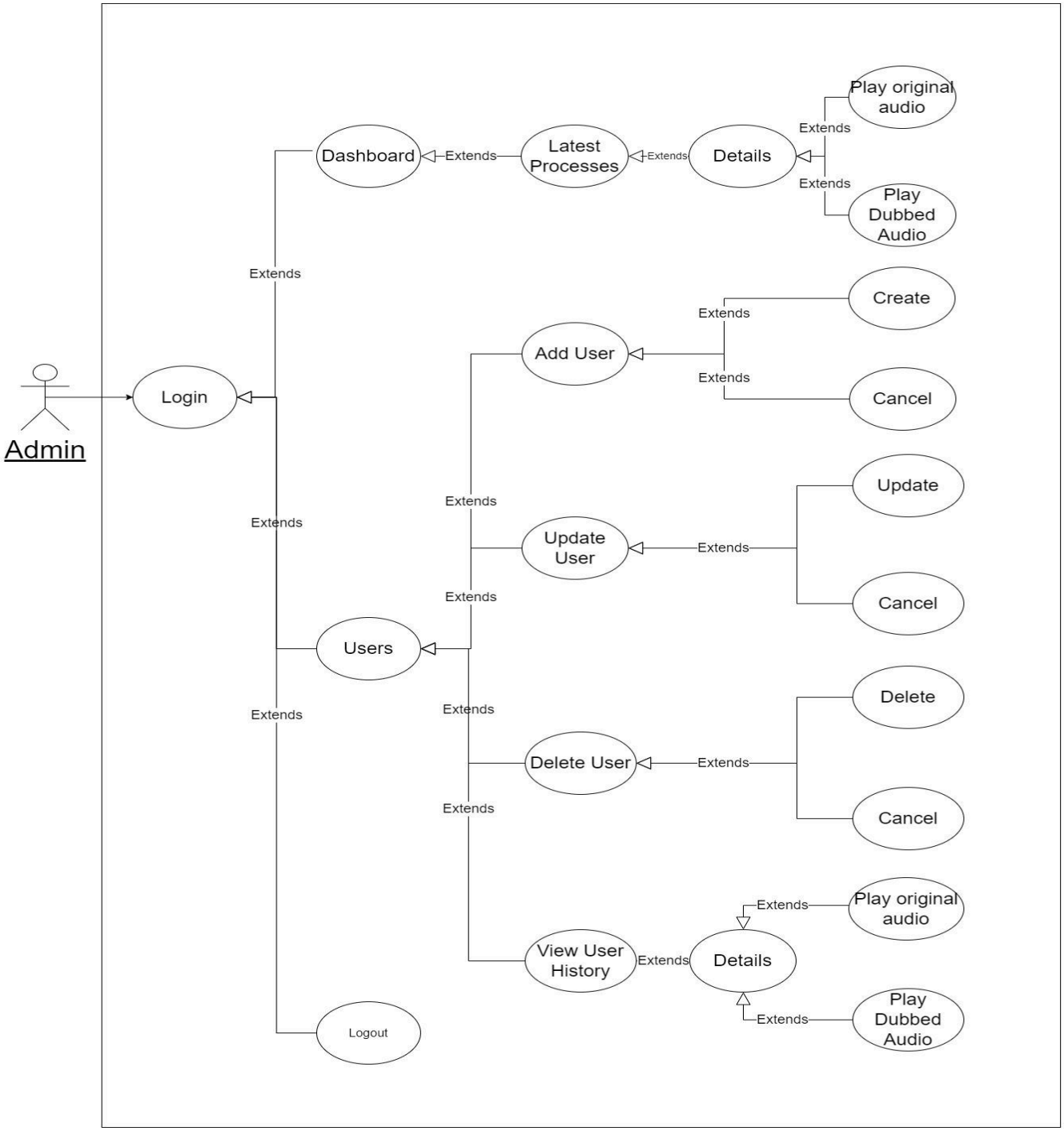


Figure [11] (Same Use Case Divided into three clear parts)

4.8.2 Admin Use Case Figure [12]:



4.9 Use Case Specification:

4.9.1 Login:

Title: Login	
Actor: User	
Description: Logs the user into the system.	
Precondition: User Created an account.	
Postcondition: Allows the user to access the application features.	
The flow of events:	
Actor:	System:
1-Fill out fields	
2-Click login	
	3-Validate Information
	4-Log the user in
	5-Show success message
Critical Scenario	
Error	System Response
Incorrect field	Highlight field
User does not exist	Show message

Figure [13]

4.9.2 Upload Audio:

Title: Upload Audio	
Actor: User	
Description: User uploads an audio file.	
Precondition: User Logged in.	
Postcondition: Listen to dubbed audio.	
The flow of events:	
Actor:	System:
1-Click on choose audio file	
2-Name audio file	
3-Click on Upload audio	
	4-Audio uploaded
	5-Show success message
Critical Scenario	
Error	System Response
Audio is not uploaded	Show message

Figure [14]

4.9.3 Create Playlist:

Title: Create Playlist	
Actor: User	
Description: User creates playlist.	
Precondition: User Logged in.	
Postcondition: Listen to dubbed organized audios.	
The flow of events:	
Actor:	System:
1-Click on add playlist	
2-Name and describe playlist	
3-Click on create playlist	
	4-Playlist Created
Critical Scenario	
Error	System Response
Fields should not be empty	Show message

Figure [15]

4.9.4 Voice Recorder

Title: Voice Recorder	
Actor: User	
Description: User records their voice .	
Precondition: User Logged in.	
Postcondition: None.	
The flow of events:	
Actor:	System:
1-Click on record audio	
	2-Fetching sound
3-Click on pause	
	4-Pauses recording
5-Click on Stop	
	6-Stops recording
	7-Shows playback, cancelation and proceeding
Critical Scenario	
Error	System Response

Figure [16]

4.9.5 Edit Profile

Title: Edit Profile	
Actor: User	
Description: User updates their profile information.	
Precondition: User Logged in.	
Postcondition: None.	
The flow of events:	
Actor:	System:
1-Update fields	
2-Click update profile picture	
3-Click save	
	4-Update successful
	5-Show success message
Critical Scenario	
Error	System Response
Username or email are taken	Show message

Figure [17]

4.9.6 Show User History

Title: Show User History	
Actor: Admin	
Description: Shows the original and dubbed history of the user.	
Precondition: Admin logged in	
Postcondition: Allows the admin to listen to audio and check the text of the original and dubbed audio.	
The flow of events:	
Actor:	System:
1-Click User History	
	2-Shows all user history
3-Click play audio	
	4-Audio plays
5-Click text	
	6-Shows text
Critical Scenario	
Error	System Response
Text does not exist	Show message

Figure [18]

4.9.7 Add New User

Title: Add New User	
Actor: Admin	
Description: Admin add a new user.	
Precondition: Admin Logged in.	
Postcondition: View, update or delete users.	
The flow of events:	
Actor:	System:
1-Click on add new user	
2-Fill the required fields	
3-Click on create user	
	4-User created
	5-Show success message
Critical Scenario	
Error	System Response
Fields should not be empty	Show message

Figure [19]

4.9.8 Latest Processes

Title: Latest Processes	
Actor: Admin	
Description: Admin views and listens to the ten latest dubbed processes .	
Precondition: Admin Logged in.	
Postcondition: Play dubbed processes.	
The flow of events:	
Actor:	System:
1-Click on latest processes	
	2-Gets latest ten dubbed processes
3-Click on play audio	
	4-Audio plays
5-Click on show text	
	6-Display text
Critical Scenario	
Error	System Response
Fields should not be empty	Show message

Figure [20]

Chapter 5: System Design

5.1 Block Diagram

5.1.1 System High-Level Block Diagram

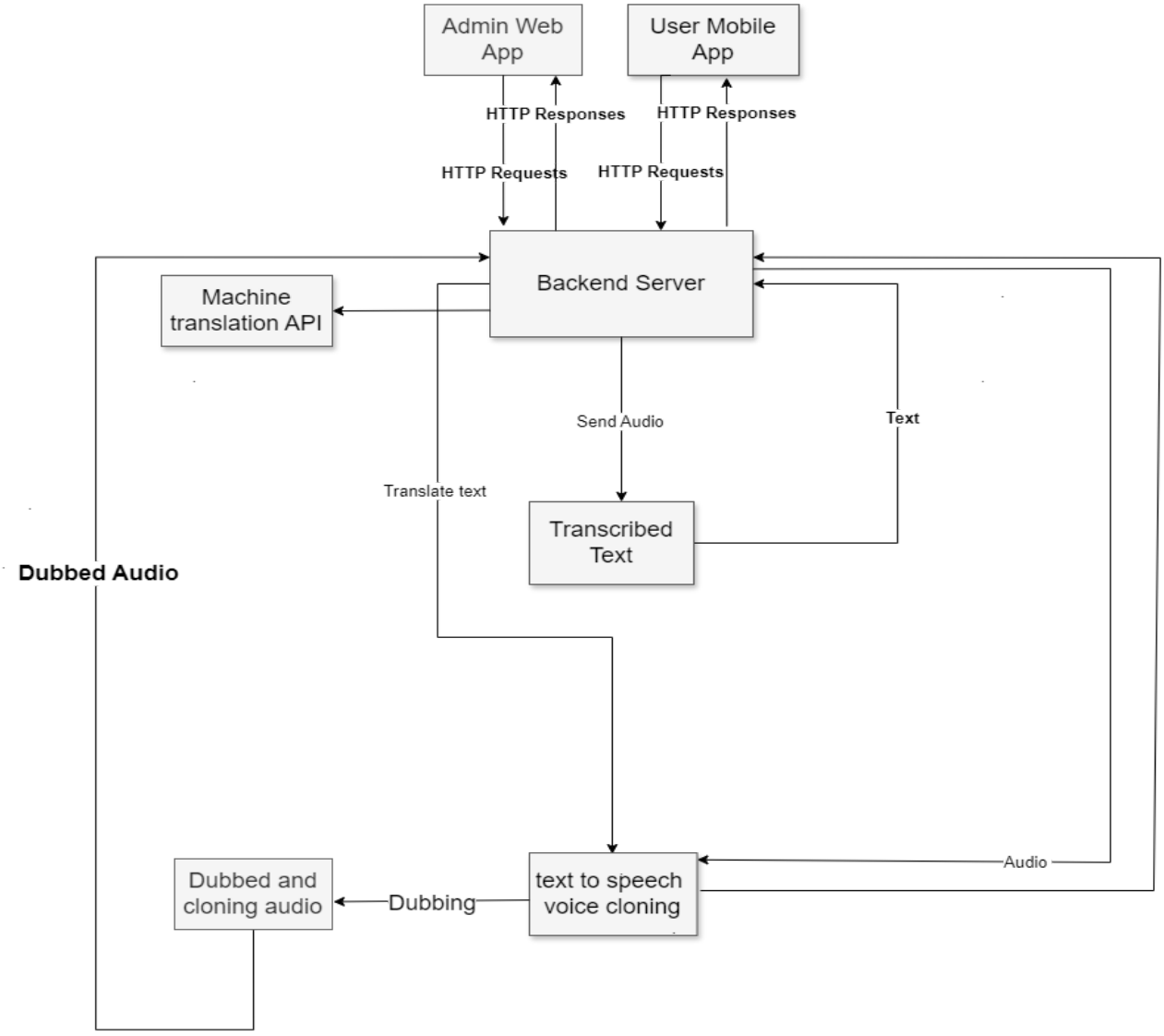


Figure [21]

5.1.2 System Low Level Block diagram

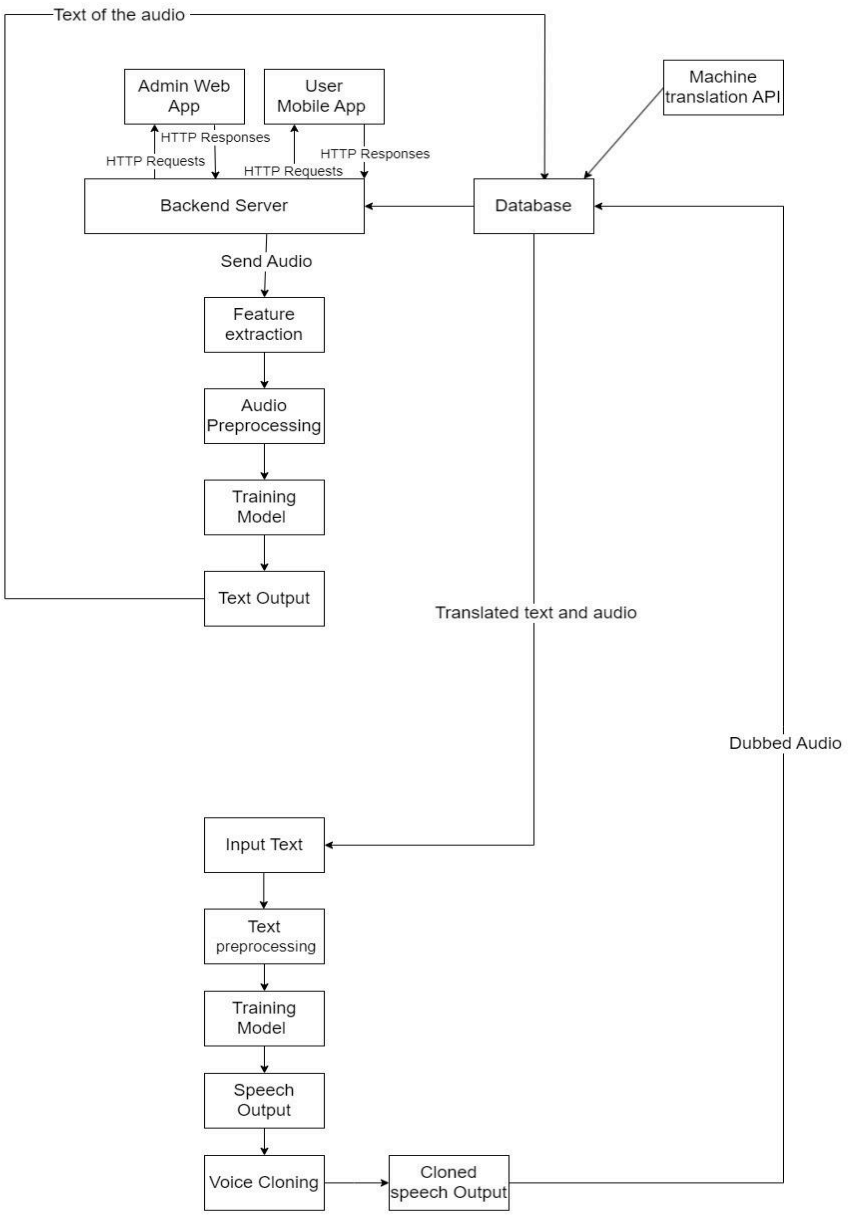


Figure [22]

5.1.3 Speech-to-Text Block Diagram

5.1.3.1 Pretrained STT Block Diagram

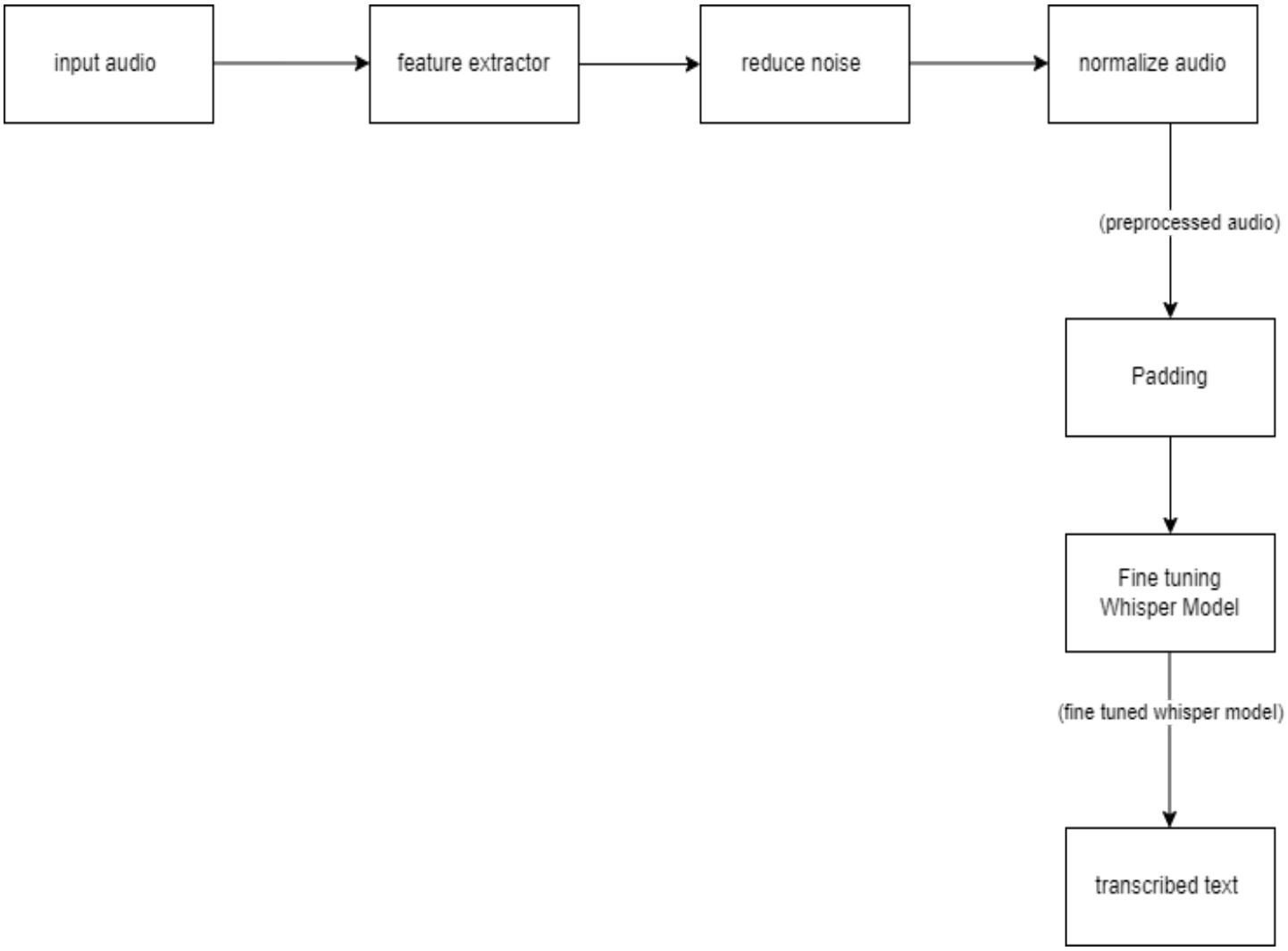


Figure [23]

5.1.3.2 From Scratch STT Block Diagram

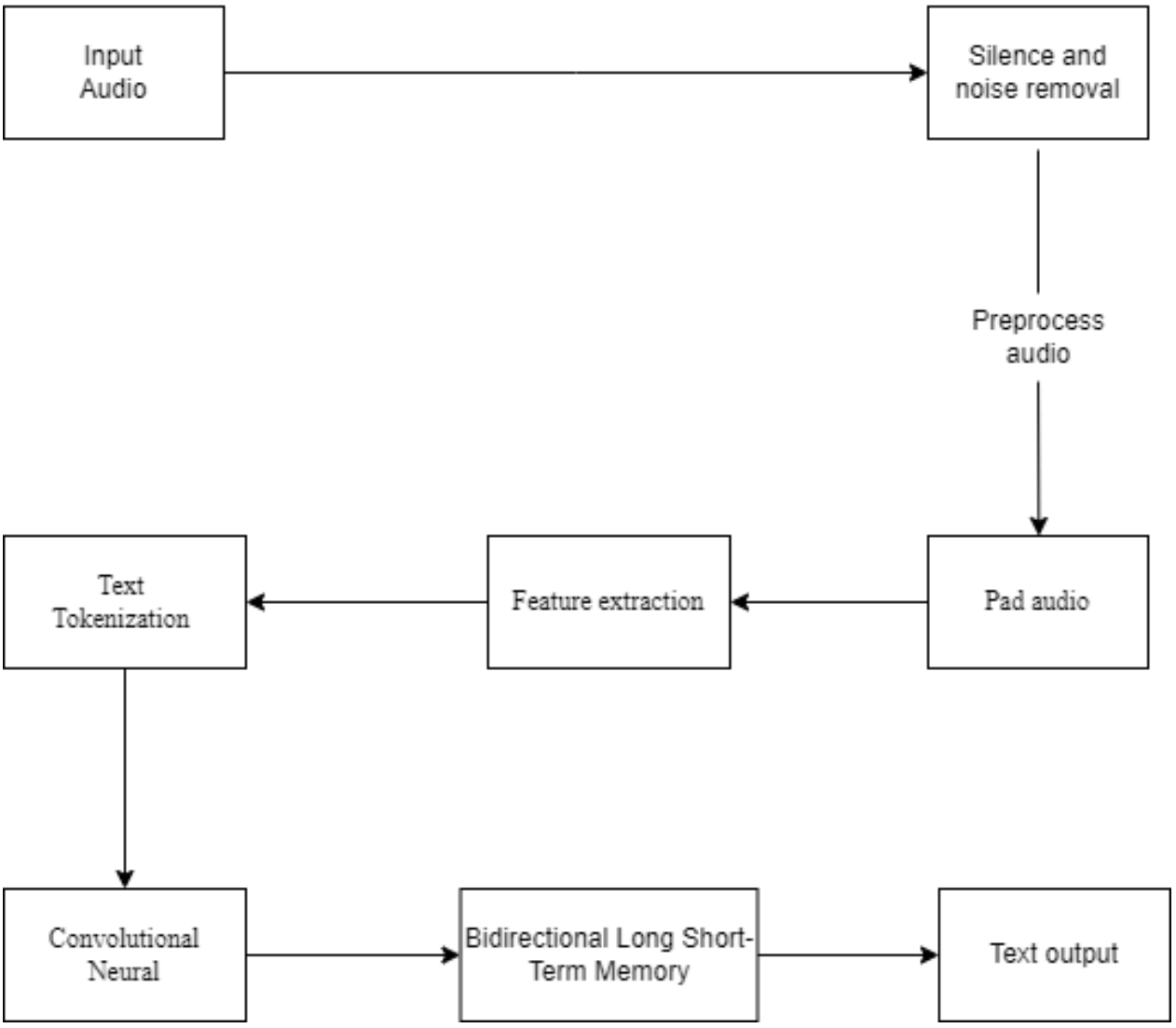


Figure [24]

5.1.4 Text-to-Speech and Voice Cloning Block Diagram

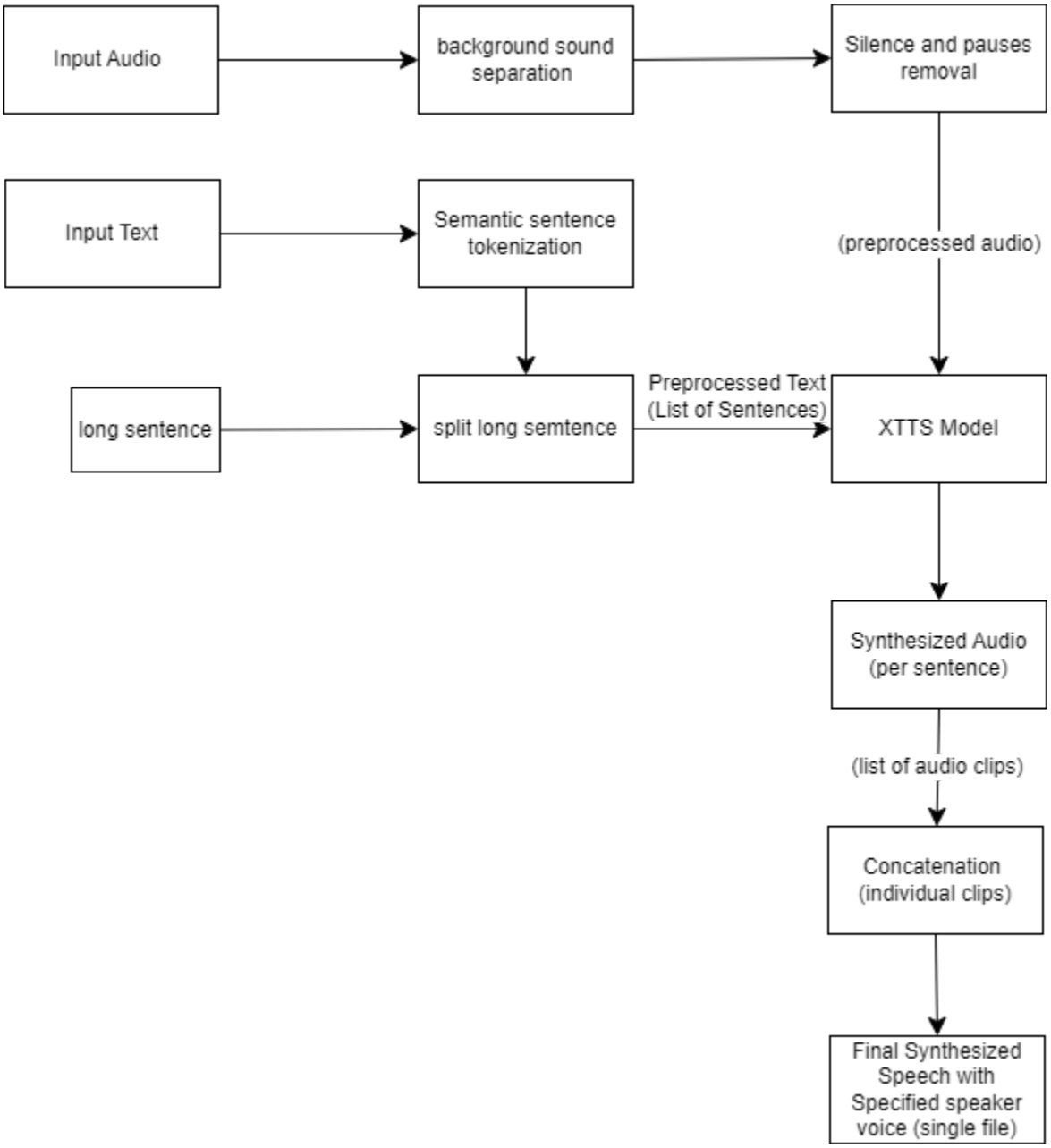


Figure [25]

5.2 ERD Diagram

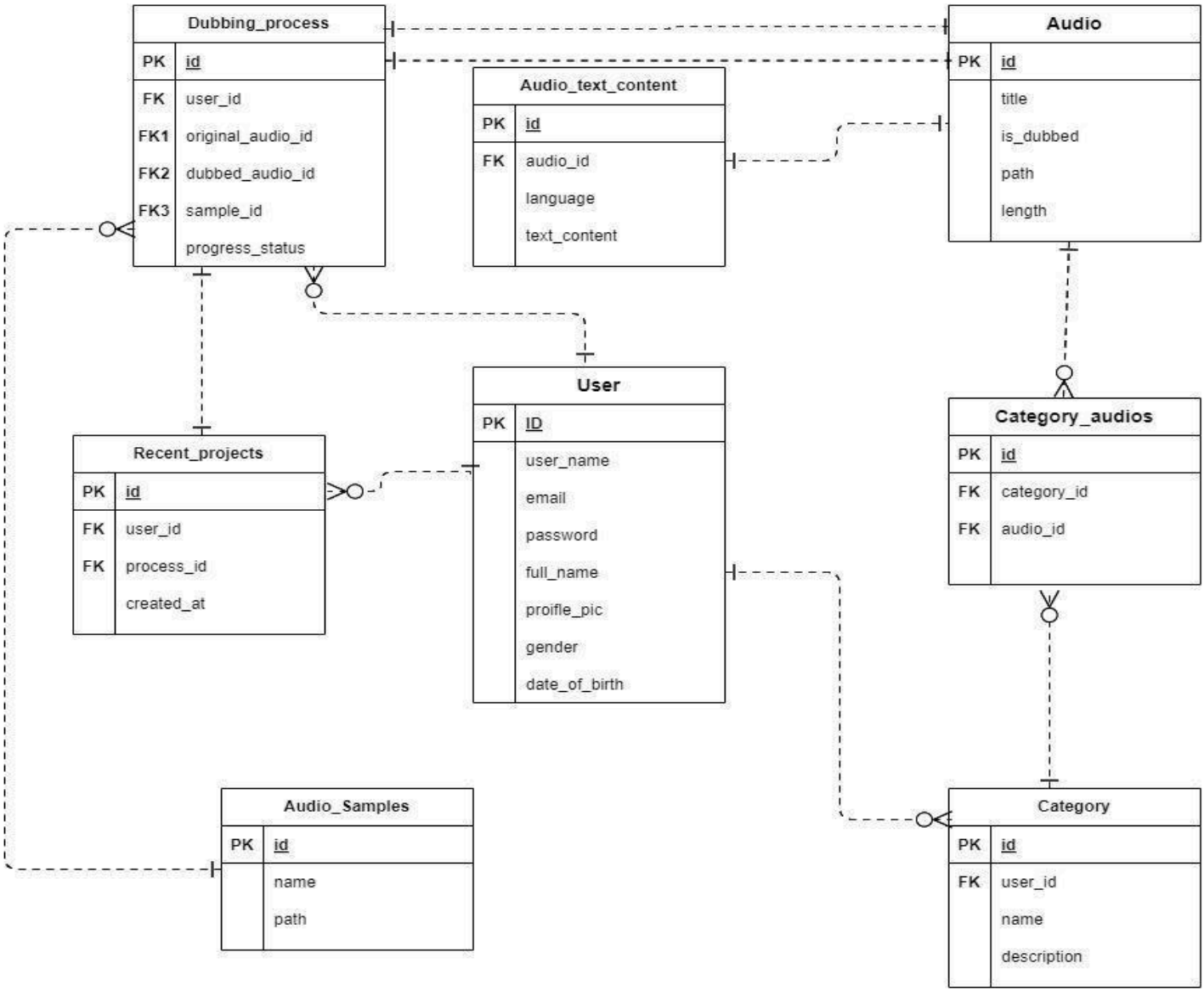


Figure [26]

5.3 Collaboration Diagram

5.3.1 User Login

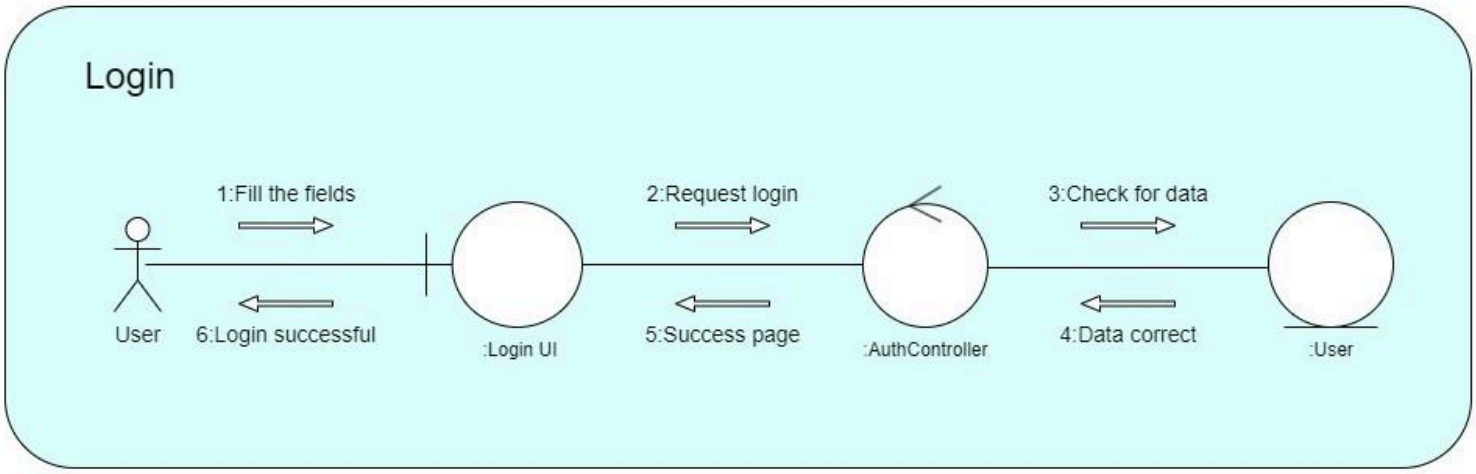


Figure [27]

5.3.2 User Listen to Dubbing

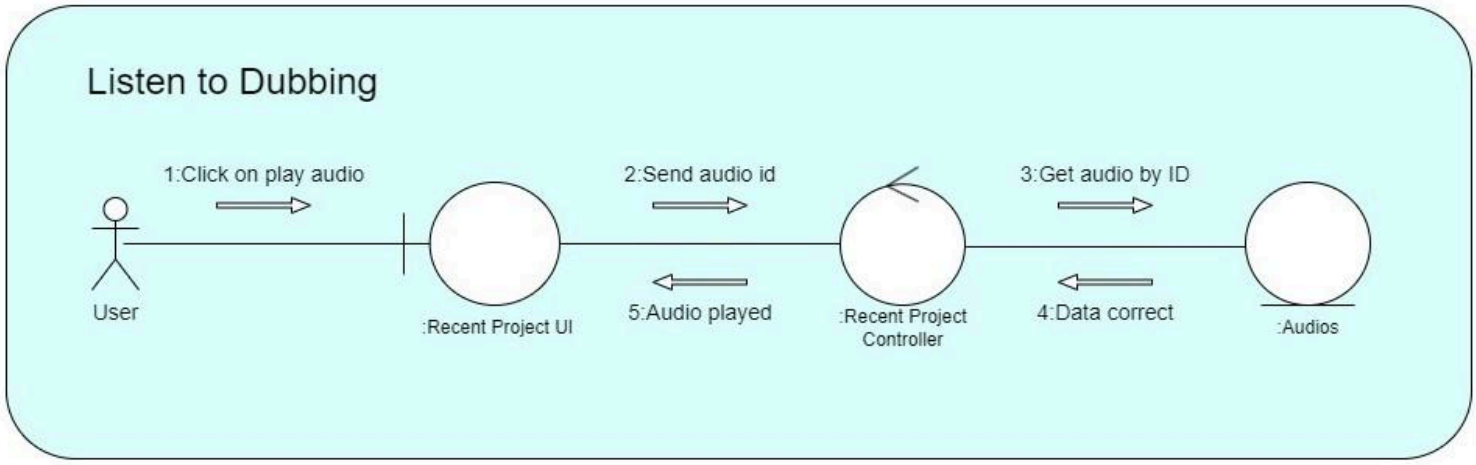


Figure [28]

5.3.3 User Delete Playlist

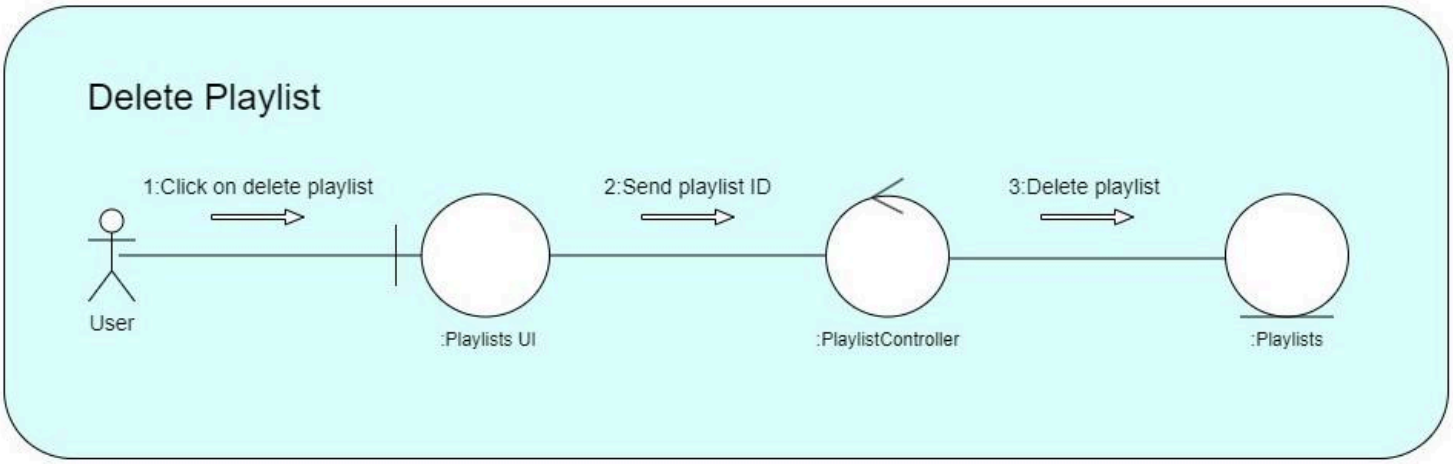


Figure [29]

5.3.4 User Record Voice

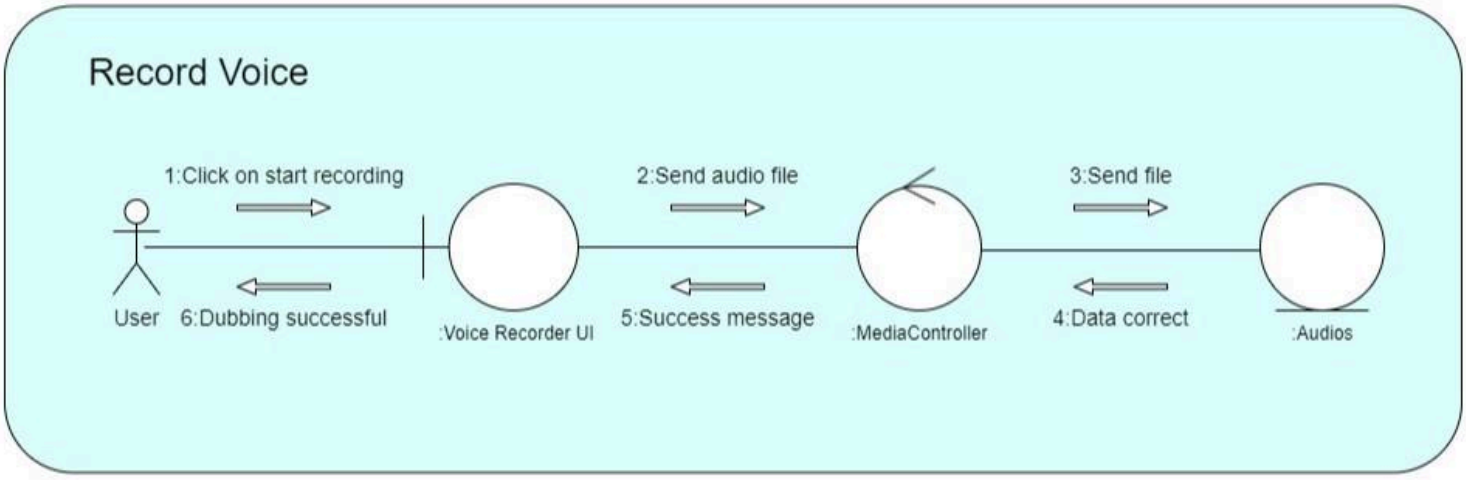


Figure [30]

5.3.5 User Upload Audio

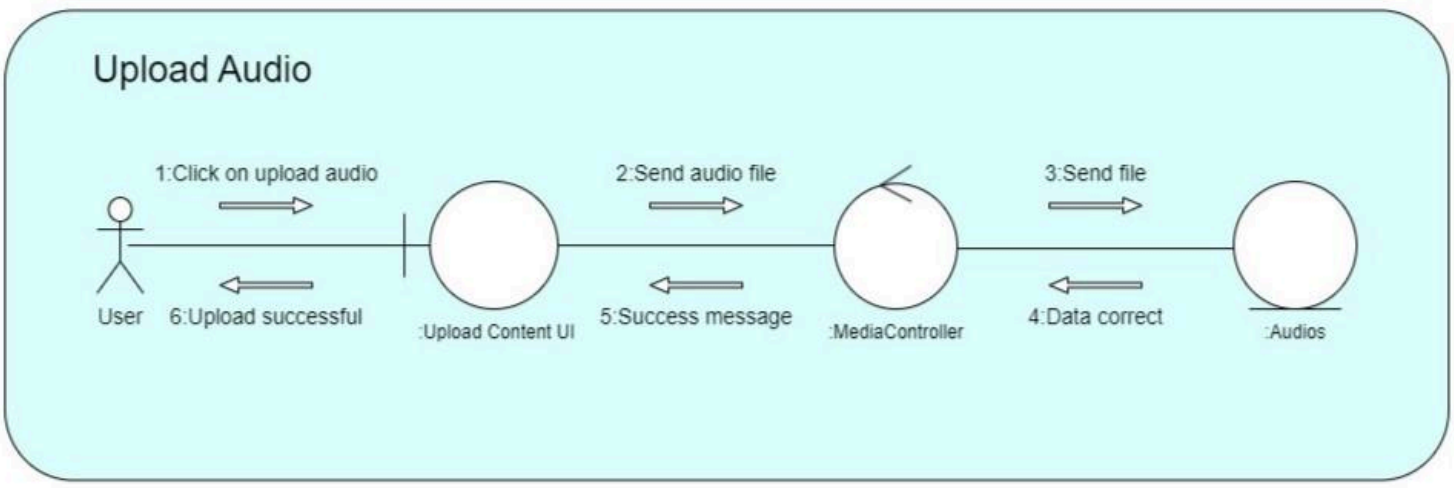


Figure [31]

5.3.6 Admin View User History

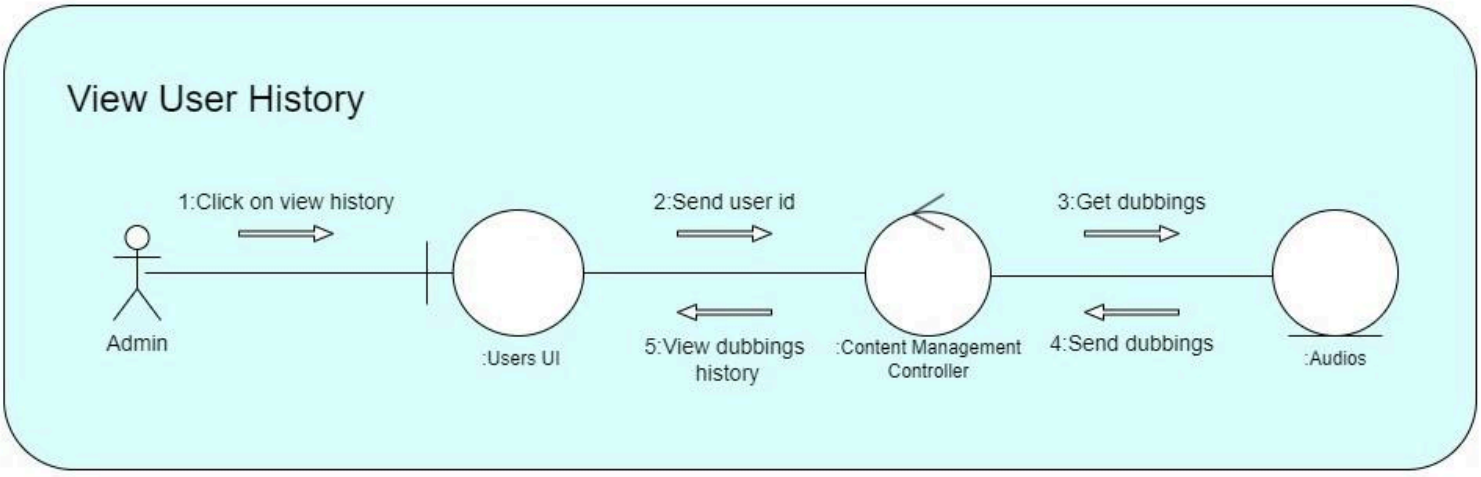


Figure [32]

5.3.7 Admin Add New User

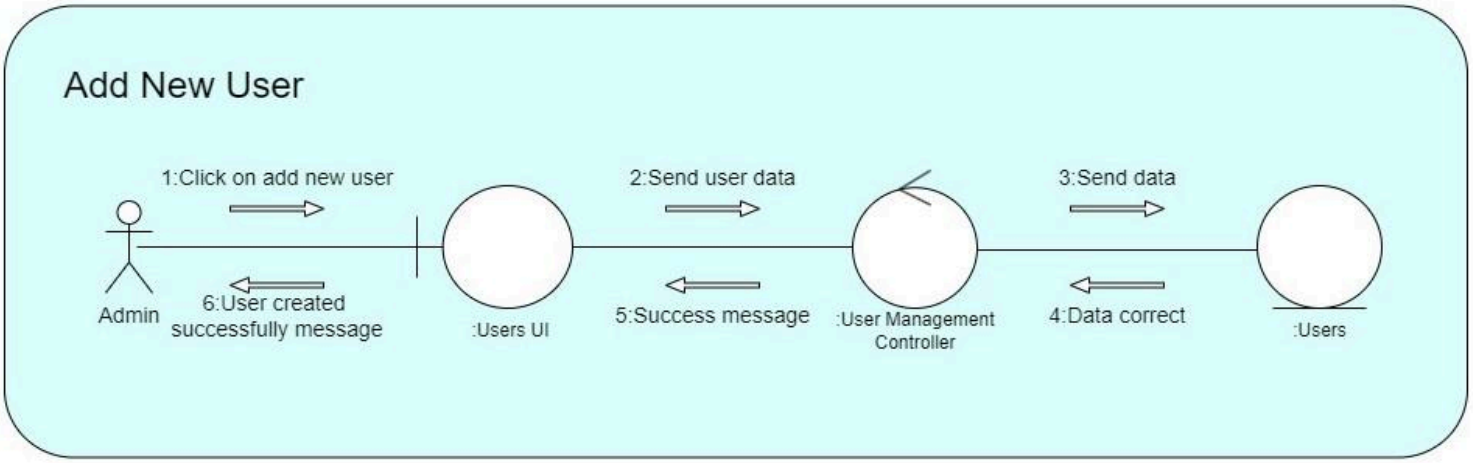


Figure [33]

5.3.8 Admin Latest Processes

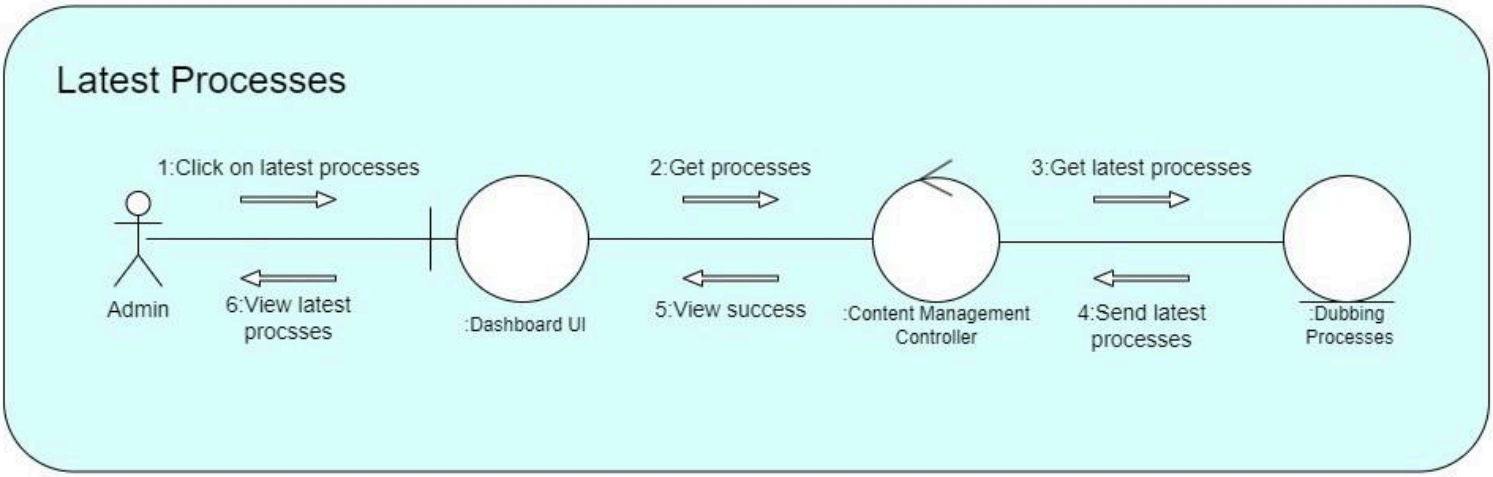


Figure [34]

5.4 Sequence Diagram

5.4.1 Upload Audio:

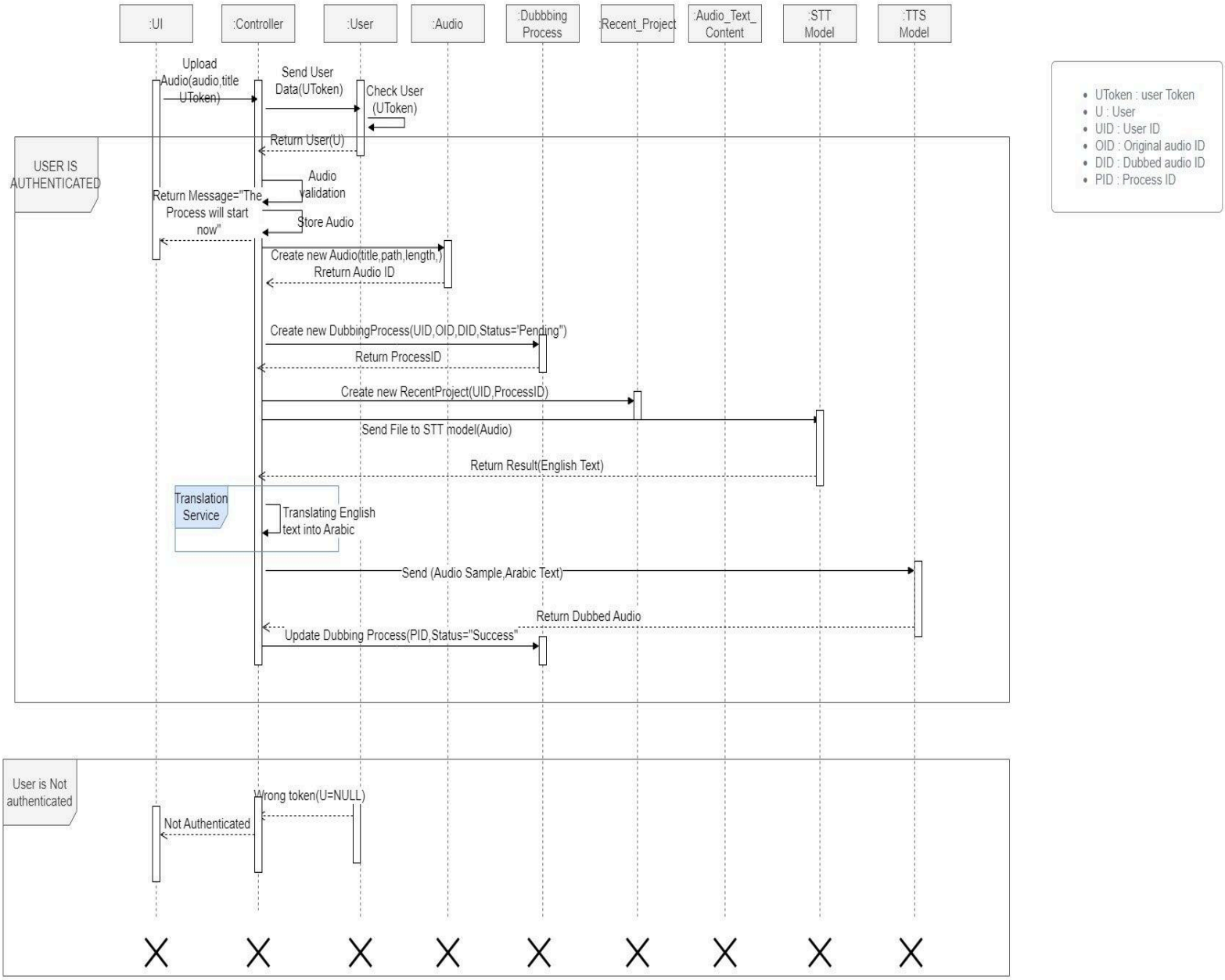


Figure [35]

5.4.2 Add Audio to Playlist:

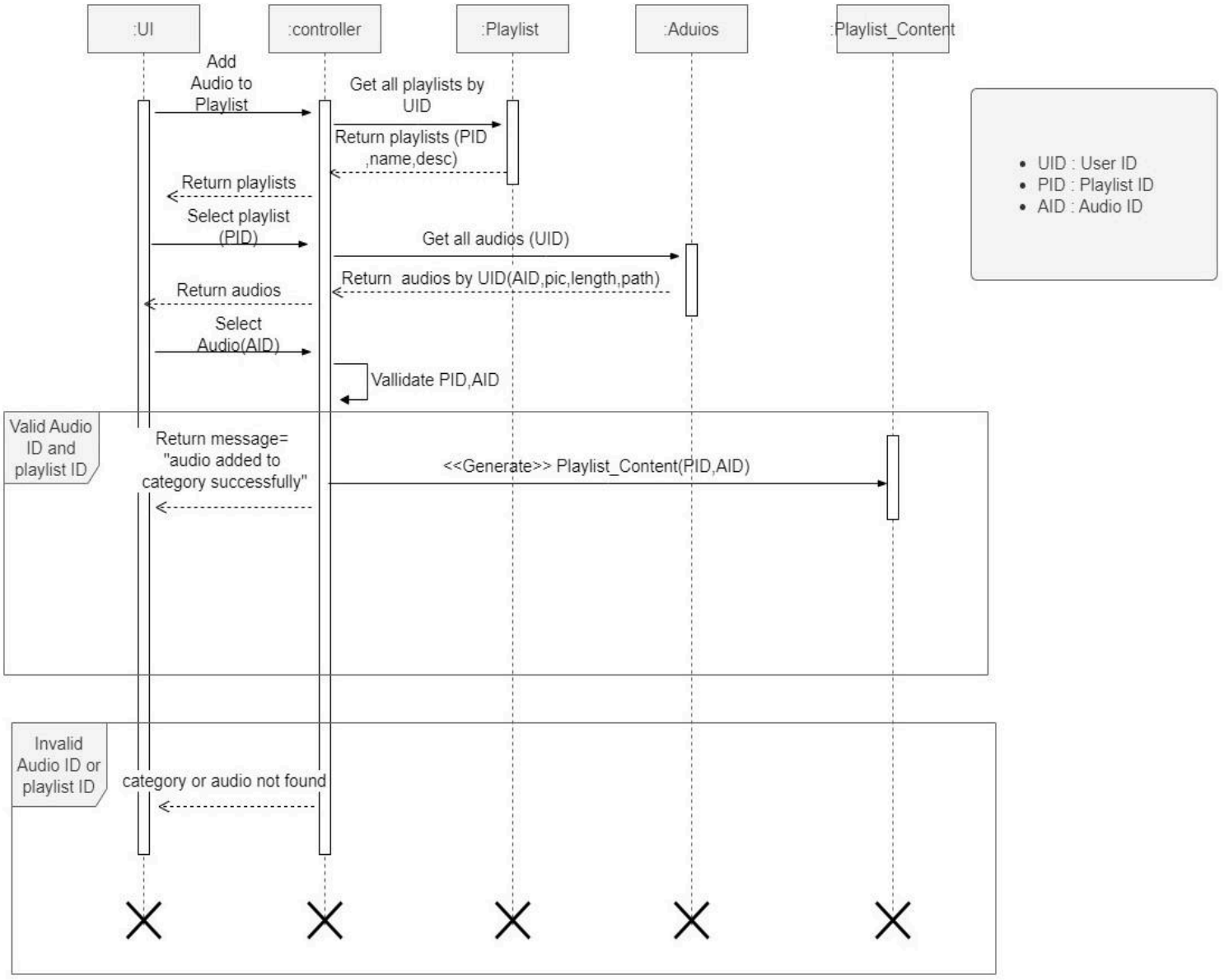


Figure [36]

5.4.2 View Recent Projects:

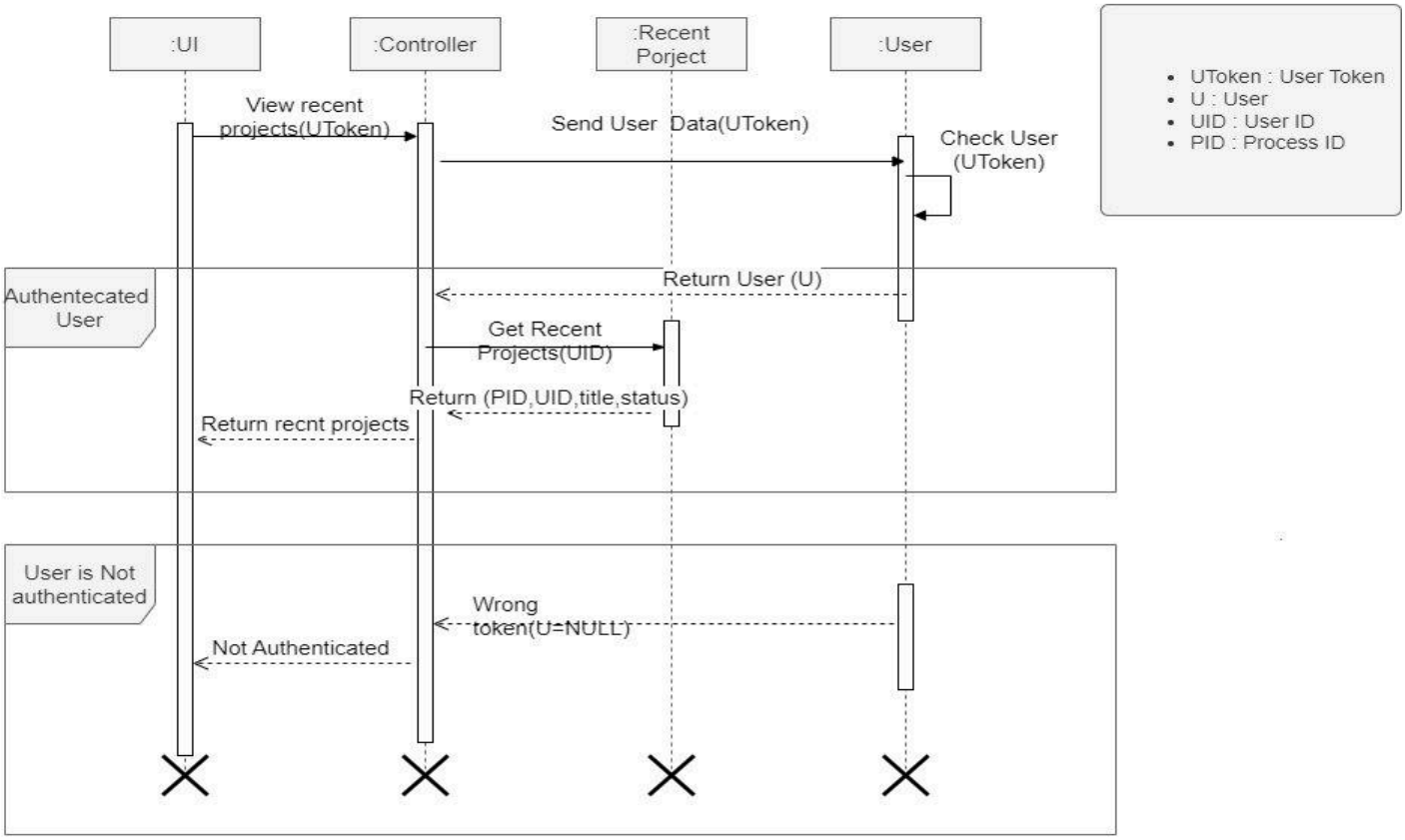


Figure [37]

5.5 AI Models

5.5.1 Speech-to-Text

5.5.1.1 STT From Scratch

Building an STT model from scratch involves various techniques, starting from preprocessing audio files to achieving accurate transcription. This process encompasses several critical steps:

- **Resampling :**The audio samples were subjected to a resampling process to standardize the sampling rate. Initially, the audio data may have been recorded at varying sampling rates. To ensure uniformity and compatibility for further processing and analysis, the audio samples were resampled to a standardized sampling rate of 16,000 Hz. This resampling process helps in maintaining consistency across different datasets and facilitates the application of various signal processing techniques under a common framework.
- **Silence and noise removal:** Unwanted silences and noise within the speaker's speech are removed before transcription.
- **Pre-emphasis:** A pre-emphasis filter is used to amplify the high-frequency parts of a signal. This is often done before transmitting or storing the signal to keep the high-frequency parts more prominent.
- **Normalization:** Audio normalization is a technique utilized in digital audio processing to adjust and standardize the volume levels of audio recordings. This process ensures a uniform loudness across multiple audio files, thereby eliminating abrupt variations in volume .

The output of this stage is a clean audio clip containing only the speaker's voice, ready for the next step which is padding .

2. Audio Padding:

This stage involves the padding of audio recordings to standardize their lengths, ensuring uniformity across all files. By extending shorter audio tracks to match the duration of the longest track, this process facilitates consistent temporal alignment, which is crucial for subsequent analytical or comparative tasks

3. Text Tokenization:

Text tokenization is employed to address the incompatibility of raw text with neural networks. This process converts text into smaller units, such as words or subwords, which are then represented as numerical vectors. Tokenization transforms the input text into a format that neural networks can effectively process and analyze, enabling more efficient and accurate text-based predictions and classifications. The output of this stage is a list of numerical tokens.

4. Feature Extraction:

Normalized Mel spectrograms are a fundamental tool in audio feature extraction, providing a visual representation that encapsulates both temporal and frequency information of an audio signal. This method transforms the audio signal into a format that is conducive to comprehensive analysis of its characteristics.

Hop Length and Overlap:

The hop length in the Short-Time Fourier Transform (STFT) is a critical parameter, representing the number of samples between the starts of consecutive frames. It can be computed as:

$$\text{hop\_length} = \text{sampling\_rate} \times \text{window\_length\_ms} / 1000$$

For instance, with a sampling rate of 16,000 Hz and a window length of 25 ms, the hop length is:

$$\text{hop\_length} = 16000 \times 25 / 1000 = 400$$

In the case of 50% overlap, the effective hop length is: hop_lengtheffective=400/2=200

The overlap, defined as the percentage of overlap between consecutive windows, results in an overlap duration of 12.5 ms:

$$\text{overlap\_duration} = \text{hop\_length} / \text{sampling\_rate} \times 1000 = 200 / 16000 \times 1000 = 12.5$$

Frequency Resolution

The frequency resolution in STFT is inversely proportional to the window length. Ensuring nfft is greater than the hop length provides adequate frequency resolution: $\Delta f = \text{sampling\_rate} / \text{n\_fft}$

For higher frequency resolution, n_fft is typically chosen as a power of 2 and greater than the hop length.

5. Model Architecture:

The input data was reshaped to include a single color channel, facilitating compatibility with the subsequent layers of the model. The architecture of the model comprises several layers of Convolutional

Neural Networks (CNNs) utilizing the Rectified Linear Unit (ReLU) activation function. Each convolutional layer is followed by a Batch Normalization (BN) layer and a Max Pooling layer to enhance performance and reduce dimensionality, respectively. Batch Normalization stabilizes and accelerates the training process by normalizing the inputs to each layer, thereby reducing internal covariate shift. Max Pooling reduces the spatial dimensions of the input, helping to retain the most prominent features and lower the computational load of the model.

Extensive experimentation was conducted to optimize the model's performance by varying the number of layers and neurons. Before passing the data into the Bidirectional Long Short-Term Memory (BiLSTM) layer, the output from the preceding layer is reshaped to match the required input dimensions of the BiLSTM. The final layer of the model is a fully connected layer with a number of neurons equal to the number of output classes. This layer employs the softmax activation function to produce a probability distribution over the output classes, facilitating accurate classification.

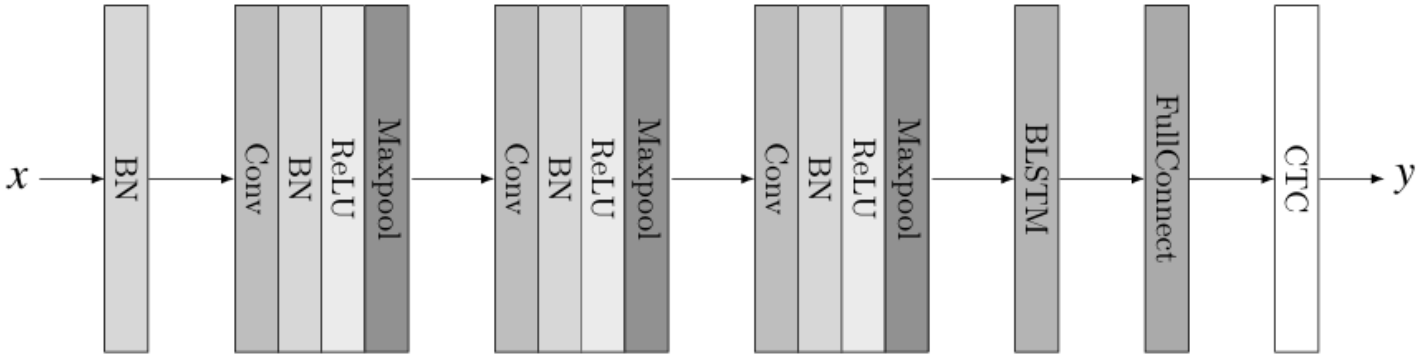


Figure [38]

5. Evaluation :

For the evaluation, Word Error Rate (WER) and Character Error Rate (CER) were utilized.

5.5.1.2 STT Pretrained

5.5.1.2.1 Whisper:

Whisper is an automatic speech recognition (ASR) system trained on 680,000 hours of multilingual and multitask supervised data collected from the web. This extensive and diverse dataset enhances the system's robustness to accents, background noise, and technical language. Additionally, Whisper supports transcription in multiple languages and translation from those languages into English. The models and inference code are being open-sourced to foster the development of useful applications and further research on robust speech processing. Whisper is a Transformer-based encoder-decoder model, also known as a sequence-to-sequence model. It maps a sequence of audio spectrogram features to a sequence of text tokens. Initially, the raw audio inputs are converted to a log-Mel spectrogram by the feature extractor. The Transformer encoder then processes the spectrogram to generate a sequence of encoder hidden states. Finally, the decoder autoregressive predicts text tokens, based on both the previous tokens and the encoder hidden states. Figure 26 summarizes the Whisper model.

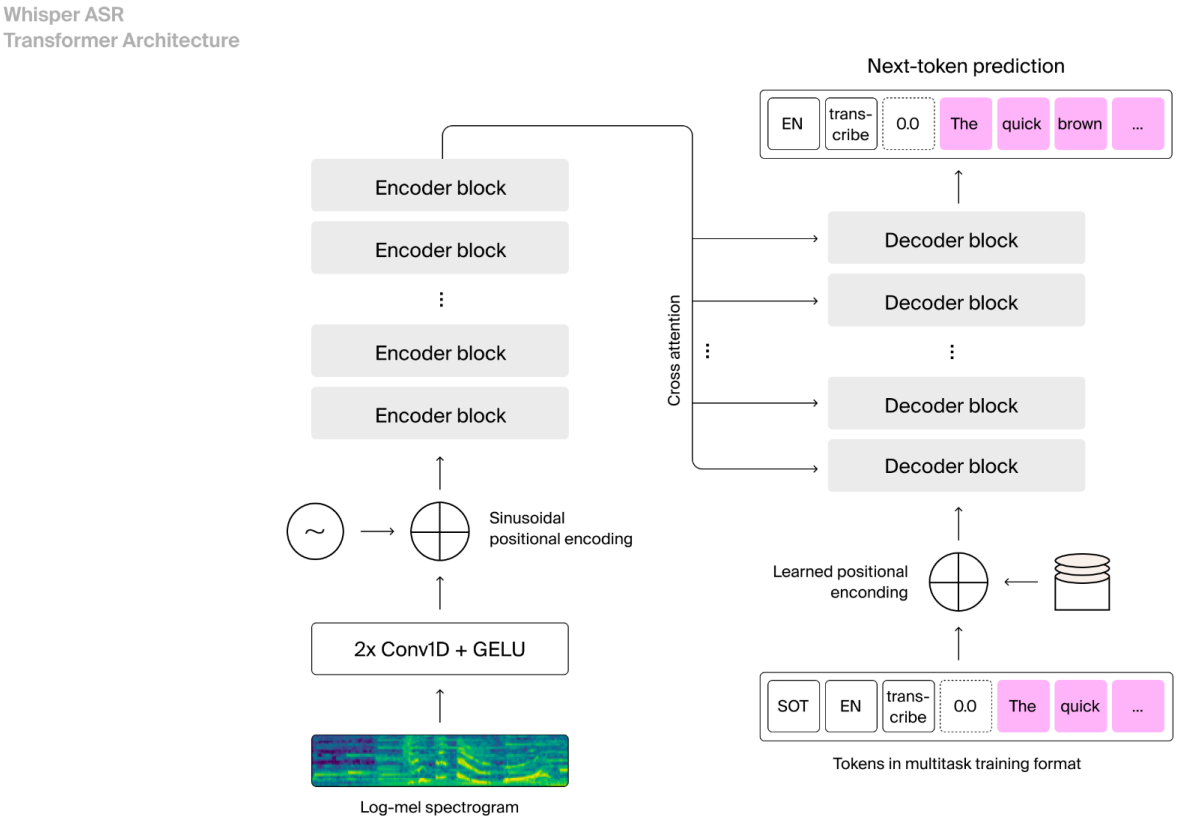


Figure [39]

5.5.1.2.2 proposed work:

Using a small whisper can serve multiple purposes, such as ensuring privacy and confidentiality, creating intimacy, and maintaining situational appropriateness. Whispering helps keep conversations discreet, prevents eavesdropping, and creates a sense of closeness. It's useful in quiet environments, like libraries or during meetings, and can emphasize the secrecy or importance of the information being shared. Whispering is also crucial for safety and security in situations where avoiding detection is necessary. Social and cultural norms often dictate when whispering is appropriate, and it can have psychological impacts like creating anticipation or providing a calming effect. Additionally, in literature

and performance art, whispering is used to convey subtlety and emotion. Overall, whispering is a powerful tool in communication that enhances privacy, intimacy, and emotional connection while being respectful of the environment and cultural norms.

5.5.1.2.3 Fine tuning Whisper model:

Model and Data Preparation:

- Loaded the Whisper model and processor from the Hugging Face library.
- Extracted and loaded the Common Voice dataset from a compressed file.
- Filtered the dataset and converted metadata into the Hugging Face dataset format.
- Split the dataset into training and testing subsets for model training and evaluation.

Data Preprocessing:

- Loaded audio files using librosa and standardized the sampling rate to match the model's requirements using the WhisperFeatureExtractor.
- Applied noise reduction techniques using librosa's noise reduction functionalities to minimize background noise in the audio files.
- Normalized the audio signals using librosa to ensure consistency in the dataset.
- Prepared the dataset by processing audio files and aligning them with corresponding text transcriptions.

Evaluation Metrics:

- Selected Word Error Rate (WER) and Character Error Rate (CER) as the evaluation metrics.
- Implemented functions to compute these metrics during training and evaluation phases.

Training Setup:

- Defined a data collator to handle padding and batching of data during training. This ensures that input sequences are appropriately padded to a uniform length, facilitating efficient batch processing.
- Created a function to compute WER and CER, ensuring that both orthographic (original text) and normalized versions are evaluated.

Training Configuration:

- Configured training parameters including batch size, learning rate, number of training steps, and other relevant hyperparameters.

Model Training:

- Initialized the Seq2SeqTrainer with the model, training and evaluation datasets, data collator, and evaluation metrics.
- Commenced the training process, periodically evaluating the model's performance using the selected metrics.

5.5.2 Text-to-Speech and Voice Cloning

Text-to-Speech and Voice Cloning builds upon XTTS, but includes several novel modifications to enable efficient Arabic TTS and improve ZS-Voice Cloning for any speaker audio the user chooses, Figure [14] shows an overview of the Text-to-Speech and Voice Cloning overall and final architecture we used to pass into our application.The model is composed of three main components:

5.5.2.1 Semantic Sentence Tokenizer:

Traditional sentence tokenization techniques typically rely on punctuation marks like full stops, question marks, and exclamation marks to segment text into sentences. While effective in many cases, this approach can lead to suboptimal results in scenarios where punctuation is missing or misused. Additionally, traditional tokenization might not always capture the true semantic meaning of a sentence, especially in complex text structures.

Semantic sentence tokenization addresses these limitations by incorporating semantic information during the segmentation process. Tools like Stanza, a natural language processing library, employ this approach. Stanza utilizes pre-trained models that analyze the context and meaning of the text to identify sentence boundaries more accurately.

- **Improved Speech Naturalness:** By ensuring sentences are split at logical points that align with meaning, semantic tokenization contributes to a more natural flow and rhythm in the synthesized speech.
- **Preservation of Context:** Accurate sentence segmentation helps XTTS maintain the intended context of the original text, leading to more faithful speech generation.
- **Enhanced Text-to-Speech Performance:** Semantic tokenization provides XTTS with a well-structured representation of the text, facilitating smoother processing and potentially improving overall speech quality.

Choosing the Sentence tokenization tool

Feature	Stanza[7]	spaCy[8]
Arabic Text Support	Pre-trained models specifically designed for Arabic	Limited native support, requires additional configuration or external models
Tokenization	Semantic, considers meaning and context	Primarily relies on punctuation
Functionality for Arabic	Out-of-the-box for tasks like POS tagging, NER, parsing	Requires additional setup and expertise
Language Support	Wider range, including Arabic	Limited native support, primarily focused on European languages

5.5.2.2 Background Sound Separation:

When working with audio recordings, particularly those captured in real-world environments, the desired speaker's voice might be obscured by background noise or music. This can significantly impact the quality of the extracted voice sample and consequently, the synthesized speech.

Spleeter, an open-source project by Deezer Research, addresses this challenge by offering a powerful background sound separation tool. It leverages pre-trained deep learning models to isolate various audio sources within a single recording. This allows for the extraction of the target speaker's voice from a mix containing background music, noise, or other interfering sounds.

- **Improved Voice Clarity:** By effectively separating the desired speaker's voice from background noise, Spleeter provides XTTS with a cleaner audio sample. This leads to more accurate voice cloning and ultimately, higher-quality synthesized speech.
- **Enhanced Text-to-Speech Performance:** A clear and isolated voice sample allows XTTS to focus on the speaker's unique vocal characteristics during the text-to-speech conversion process. This can potentially improve the accuracy and naturalness of the generated speech.
- **Flexibility in Audio Input:** Spleeter enables XTTS to handle a wider range of audio input scenarios. Even if the provided audio contains background noise or music, Spleeter can potentially extract a usable voice sample for subsequent processing

5.5.2.3 Diacritization:

This module utilizes machine learning models trained on large datasets of text with corresponding diacritics. The model analyzes the input text and attempts to predict the appropriate placement of diacritics based on various factors, including:

- **Enhanced Speech Accuracy:** By providing a more accurate representation of the intended pronunciation with diacritics, XTTS can generate speech that better reflects the nuances of the source text. This leads to improved speech quality and intelligibility.
- **Reduced Ambiguity:** Diacritics can help resolve ambiguities in text, especially in languages where pronunciation can vary depending on the presence or absence of diacritics. This ensures that XTTS interprets the text correctly, leading to more accurate speech generation.
- **Broader Text Compatibility:** The diacritization module allows XTTS to handle a wider range of text inputs, even those lacking diacritics. This improves the system's flexibility and usability.

5.5.2.4 Model Methodology

This section details the preprocessing stages that prepare the audio and text inputs for XTTS, the text-to-speech and voice cloning model.

5.5.2.4.1 Audio Preprocessing:

This step ensures XTTS receives a suitable audio input for voice cloning by handling audio clips exceeding 6 seconds. Here's how it works:

- **Splitting unwanted content:** Tools like Spleeter separate the audio from background music or noise, leaving only the speaker's voice.

- **Silence & Pause Removal:** Unwanted silence and pauses within the speaker's speech are removed to create a clean audio segment.
- **6-Second Clip Extraction:** A 6-second clip is extracted from the processed audio. This clip should ideally capture a representative sample of the speaker's voice.

The output of this stage is a clean, 6-second audio clip containing only the speaker's voice, ready for XTTS to use in voice cloning.

5.5.2.4.2 Text Preprocessing:

This stage prepares the translated text output received from the Speech-to-Text (STT) model. Here's what's involved:

1. Semantic Sentence Tokenization:

- A semantic tokenizer (tools like Stanza) segments the text into individual sentences. This ensures sentences are split at logical points, preserving the meaning and context of the original text. XTTS relies on this structure to accurately generate speech.

2. Diacritization:

- In some languages, diacritics (vowel markings) play a crucial role in pronunciation and inflection. The text is processed using a **dedicated diacritization module**. This module utilizes machine learning models trained on large datasets of text with corresponding diacritics. The model analyzes the text and attempts to predict the appropriate placement of diacritics based on various factors, including:
    - Language-specific rules

    - Contextual clues from surrounding words and sentence structure
    - Statistical probabilities from the training data
- By providing a more accurate representation of the intended pronunciation with diacritics, XTTS can generate speech that better reflects the nuances of the source text.

3. Long Sentence Handling:

- In rare cases, the tokenizer might generate a single sentence exceeding the 400-character limit imposed by XTTS. To address this, the system strategically splits the long sentence:
    - The first part becomes a separate sentence with a maximum length of 400 characters.
    - The remaining text forms a new sentence for processing.

The output of this stage is a sequence of individual sentences, each adhering to the 400-character limit, preserving the original meaning, and having diacritics added for accurate pronunciation representation. This prepared text is then ready for XTTS to perform text-to-speech conversion.

5.5.2.4.3 XTTS Model Usage:

Once both audio and text are preprocessed, they are fed into the XTTS model. Here's the process:

- **Model Input:** The preprocessed audio clip and the list of sentences are provided as inputs to XTTS.
- **Text-to-Speech Conversion (Per Sentence):** XTTS performs text-to-speech conversion on each sentence individually, leveraging the speaker's voice characteristics captured in the audio clip.
- **Audio Output (Per Sentence):** For each sentence, XTTS generates a new audio clip containing the synthesized speech.

The output of this stage is a list of individual audio clips, each corresponding to a sentence from the preprocessed text.

5.5.2.4.4 Audio Post Processing:

The final step involves combining the individual audio clips generated by XTTS:

- **Audio List Concatenation:** A suitable audio processing library (e.g., pydub) is used to concatenate the list of audio clips into a single audio file. This final audio file represents the complete synthesized speech for the entire text, maintaining the order and context of the sentences.

The final output of the entire process is a single audio file containing the synthesized speech for the entire translated text, mimicking the speaker's voice from the provided audio.

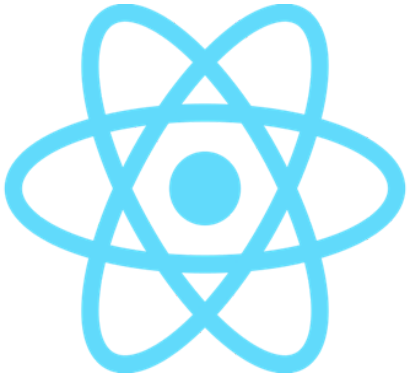
Chapter 6: Used Environments

6.1. Frontend:

6.1.1. Flutter (Dart Language):

Flutter is a UI framework that allows developers to create cross-platform mobile, web, and desktop applications using a single codebase. It leverages Dart as its primary language. Flutter provides a rich set of widgets, tools, and libraries for creating beautiful and performant user interfaces.[30]

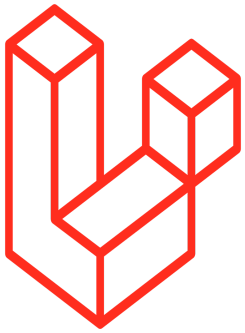
6.1.2. React JS:



React is a popular JavaScript library for building user interfaces, particularly for single-page applications. Developed and maintained by Facebook, React follows a component-based architecture, enabling the creation of reusable UI elements. Its virtual DOM efficiently updates only the necessary parts of a webpage, optimizing performance. React promotes a declarative approach to programming, making it easier to understand and debug code. With a vast ecosystem and strong community support, React is widely used for building interactive and dynamic web applications.[29]

6.2 Backend

6.2.1 Laravel



Laravel is a free and open-source PHP web application framework created
by Taylor Otwell and released in 2011. In essence, Laravel enables developers to build robust and scalable web applications using a unified and
expressive syntax. With Laravel, developers can streamline the development
process by providing built-in features for authentication, routing, sessions, caching, and database management.[27]

6.2.2 Postman

Postman is an API client that makes it easy for developers to create, share, test and document APIs. This is done by allowing users to create and save simple and complex HTTP/s requests, as well as read their responses. The result - more efficient and less tedious work.[28]

6.3 Database

6.3.1 MySQL

MySQL is an open-source relational database management system (RDBMS) Its name is a combination of "My", the name of co-founder Michael Widenius's daughter My and "SQL", the acronym for Structured Query Language.

A relational database organizes data into one or more data tables in which data may be related to each other; these relations help structure the data. SQL is a language that programmers use to create, modify and extract data from the relational database, as well as control user access to the database. In addition to relational databases and SQL, an RDBMS like MySQL works with an operating system to implement a relational database in a computer's storage system, manages users, allows for network access and facilitates testing database integrity and creation of backups.[31]

6.4 AI

6.4.1 Google Colaboratory

Colaboratory, or “Colab” for short, is a product from Google Research. Colab allows anybody to write and execute arbitrary python code through the browser, and is especially well suited to machine learning, data analysis and education. More technically, Colab is a hosted Jupyter notebook service that requires no setup to use, while providing access free of charge to computing resources including GPUs.[32]

6.4.2 PyCharm

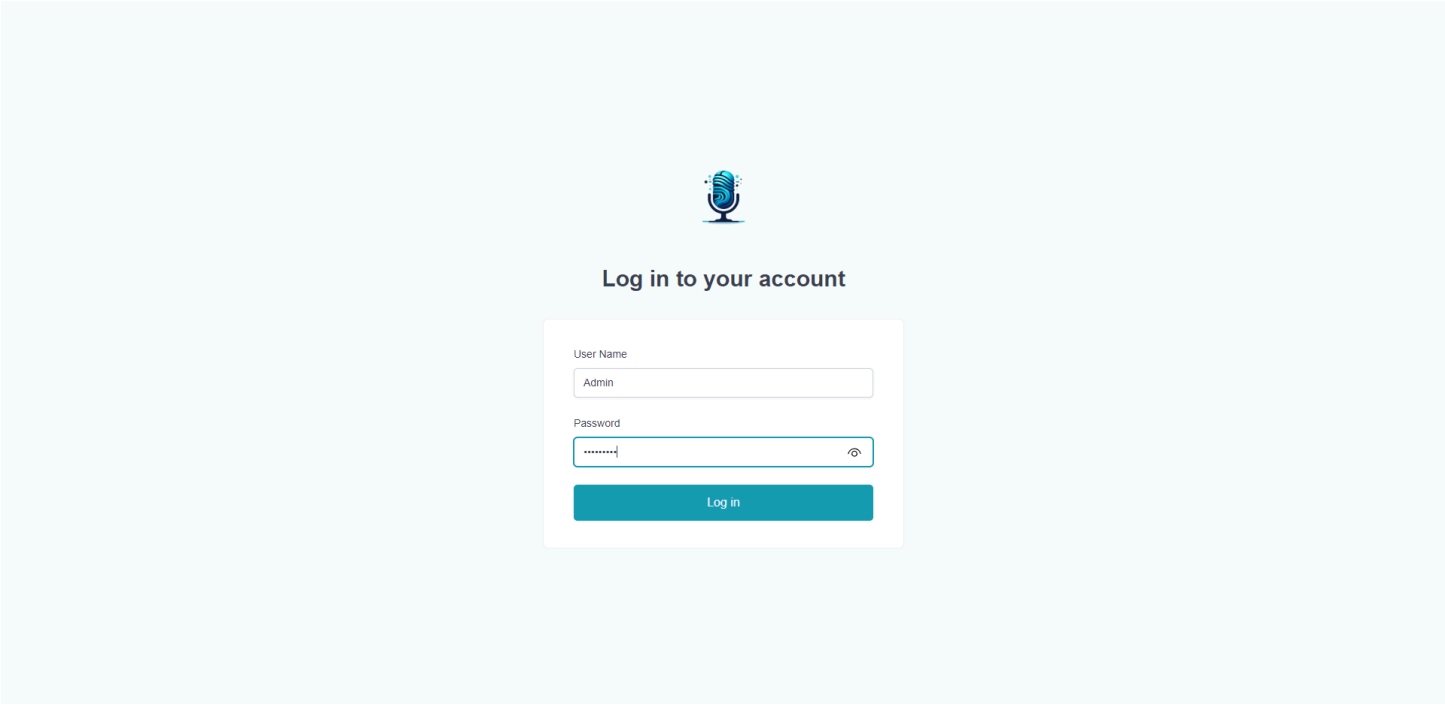
PyCharm is an integrated development environment (IDE) used for programming in Python. It provides code analysis, a graphical debugger, an integrated unit tester, integration with version control systems, and supports web development with Django. PyCharm is developed by the Czech company JetBrains.[4]

It is cross-platform, working on Microsoft Windows, macOS, and Linux. PyCharm has a Professional Edition, released under a proprietary license and a Community Edition released under the Apache License.[5] PyCharm Community Edition is less extensive than the Professional Edition.[33]

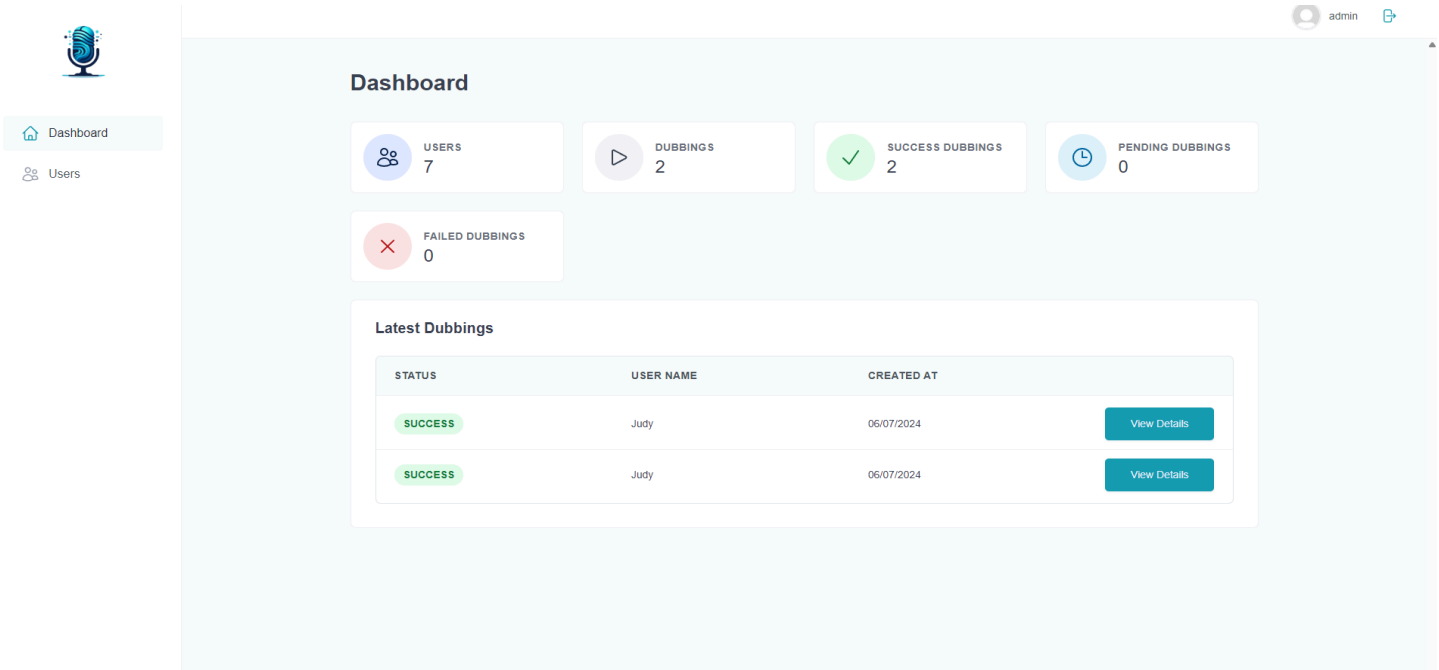
Chapter 7: Implementation

7.1. Admin Application

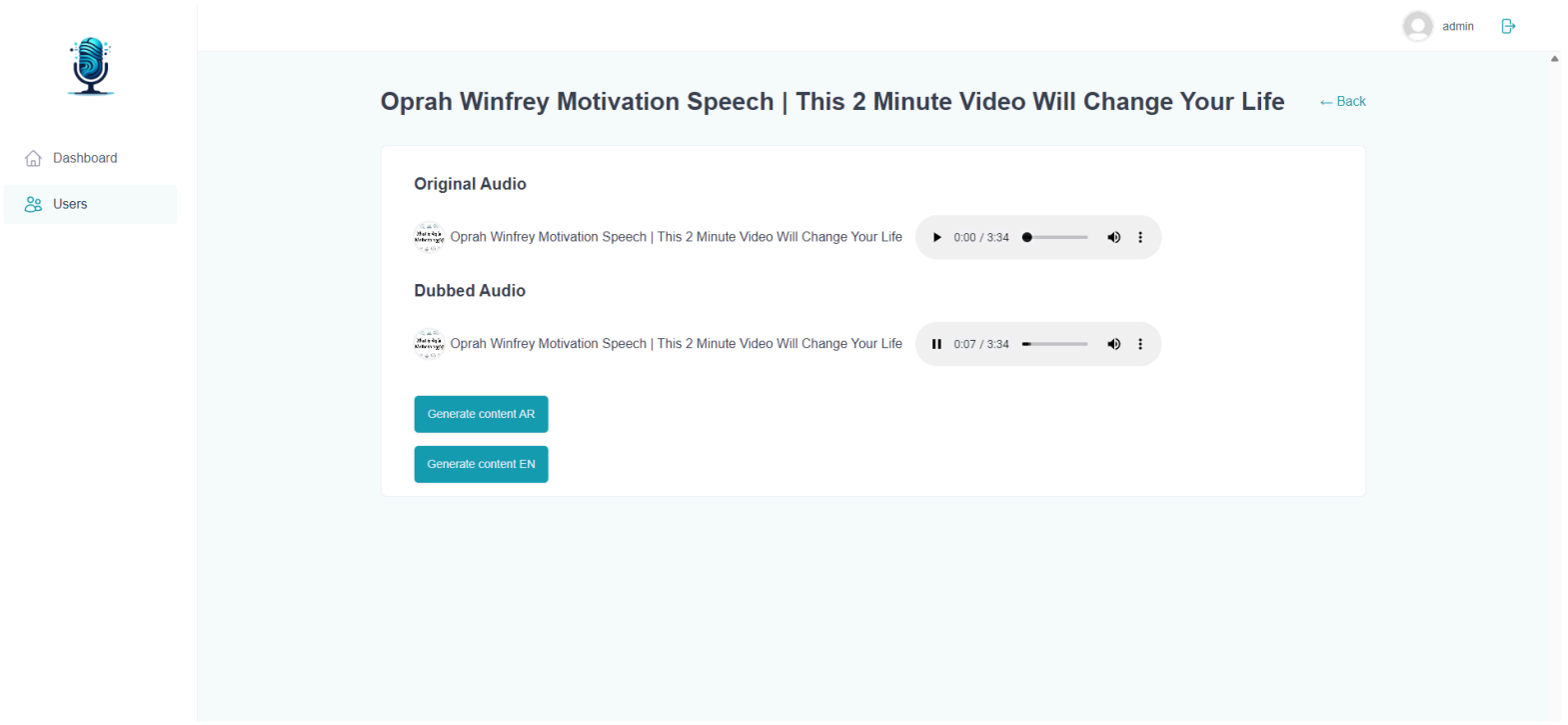
7.1.1. Login Interface: *Figure [40]*

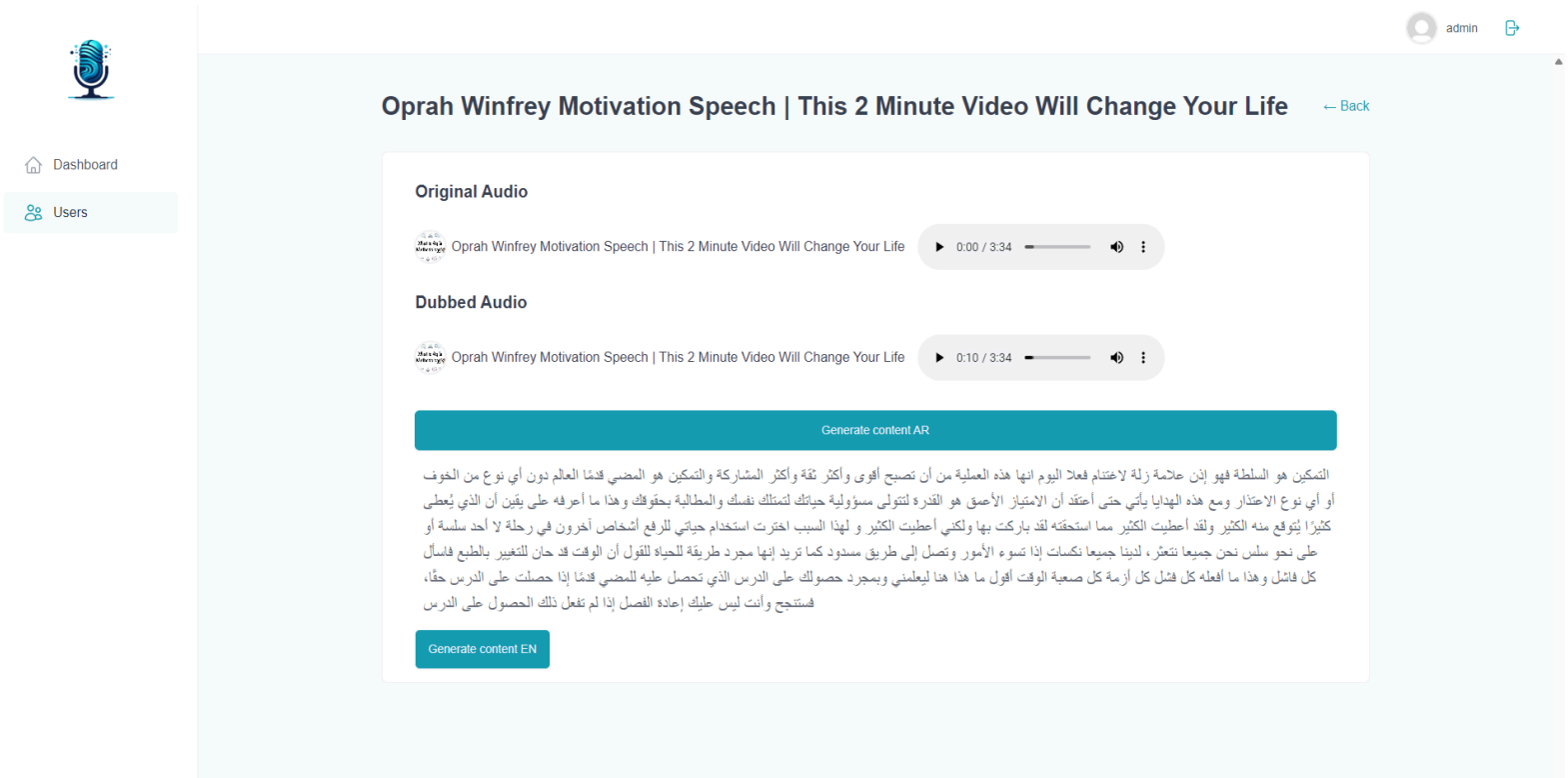


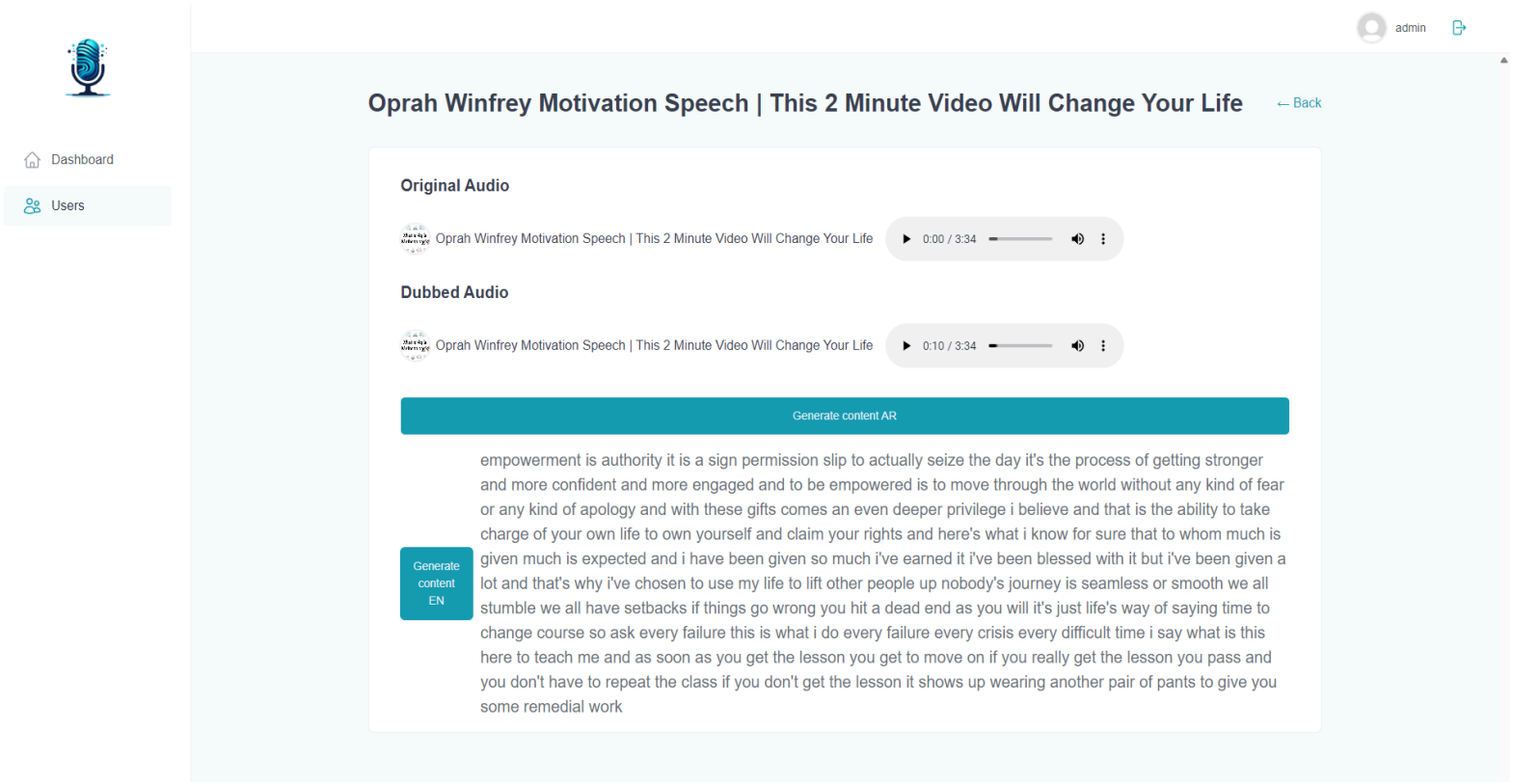
7.1.2. Dashboard Interface: *Figure [41]*



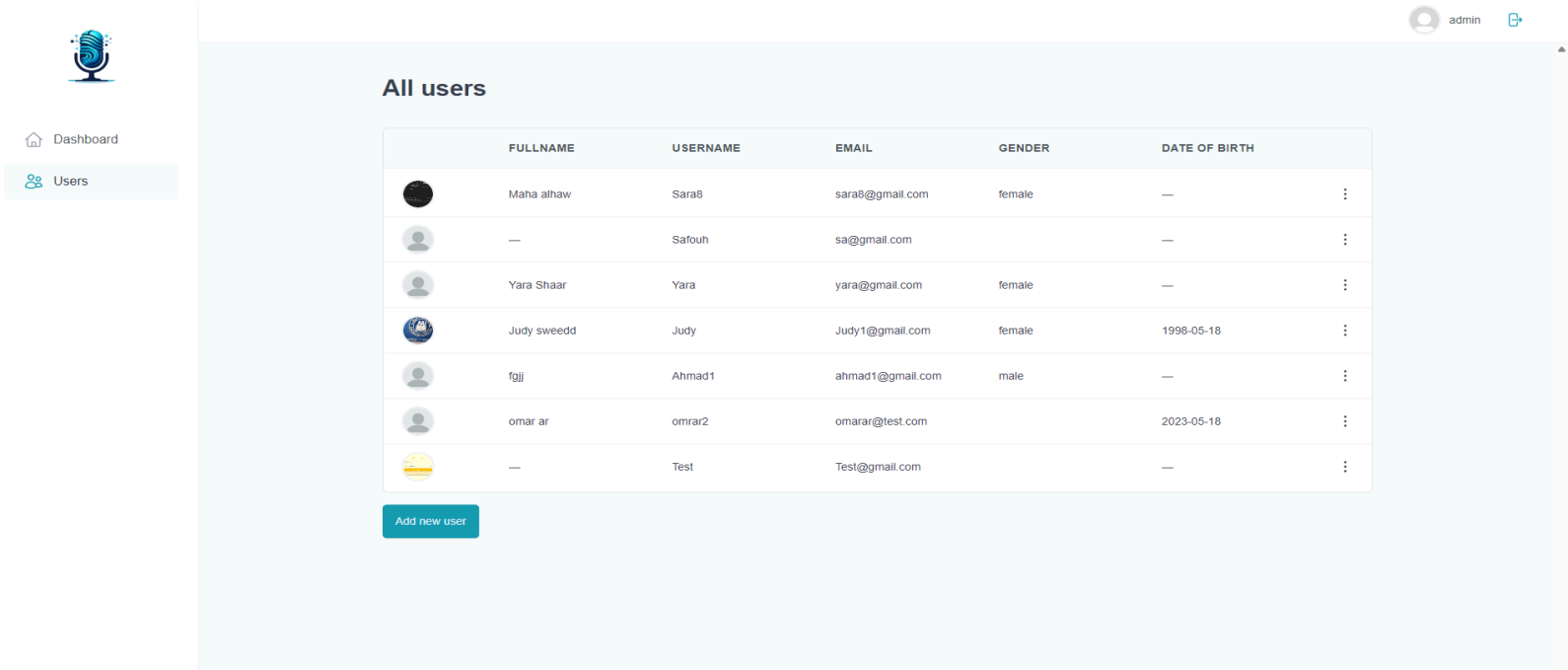
7.1.3. View Details UI: *Figure [42-43-44]*

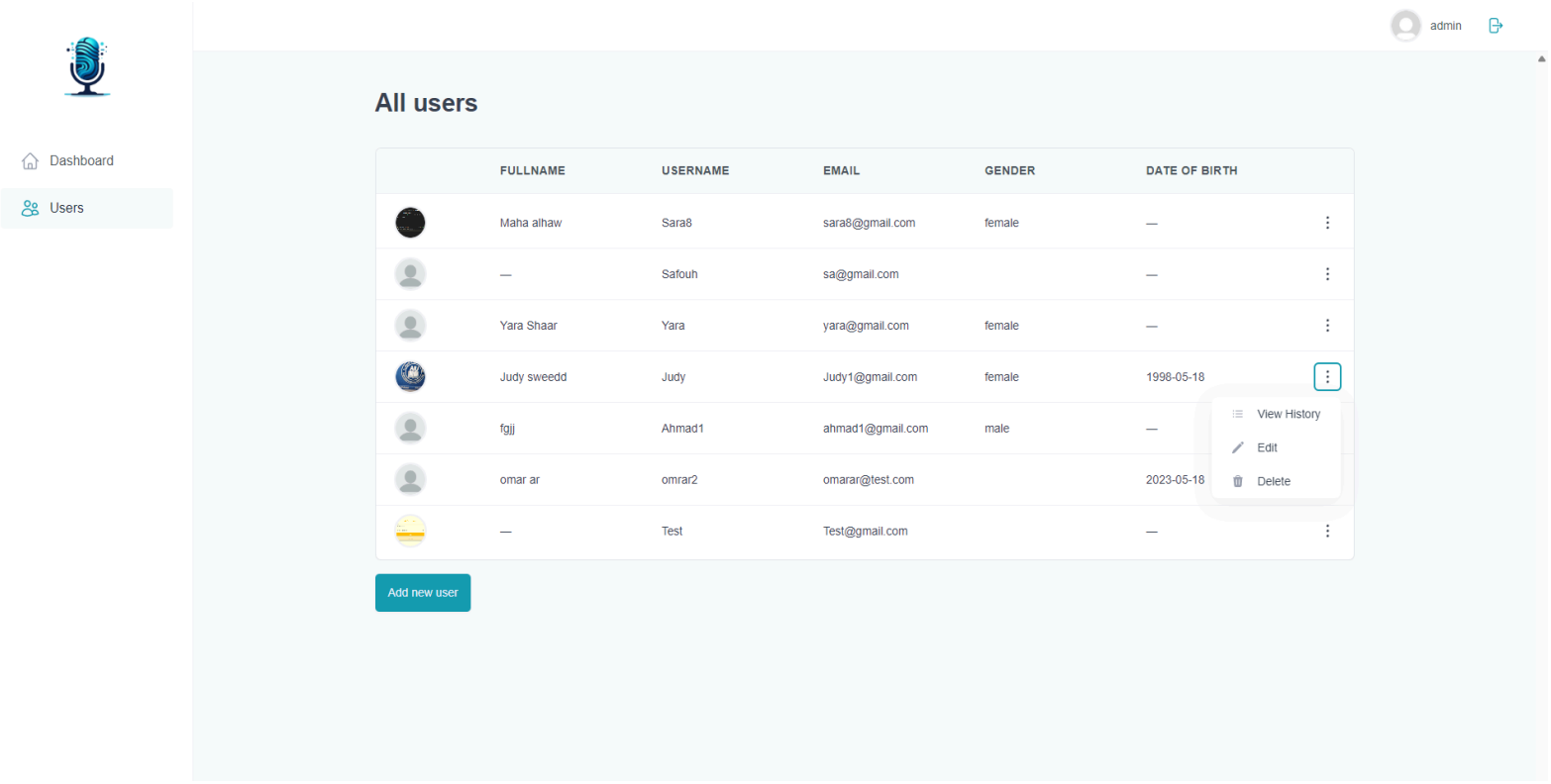




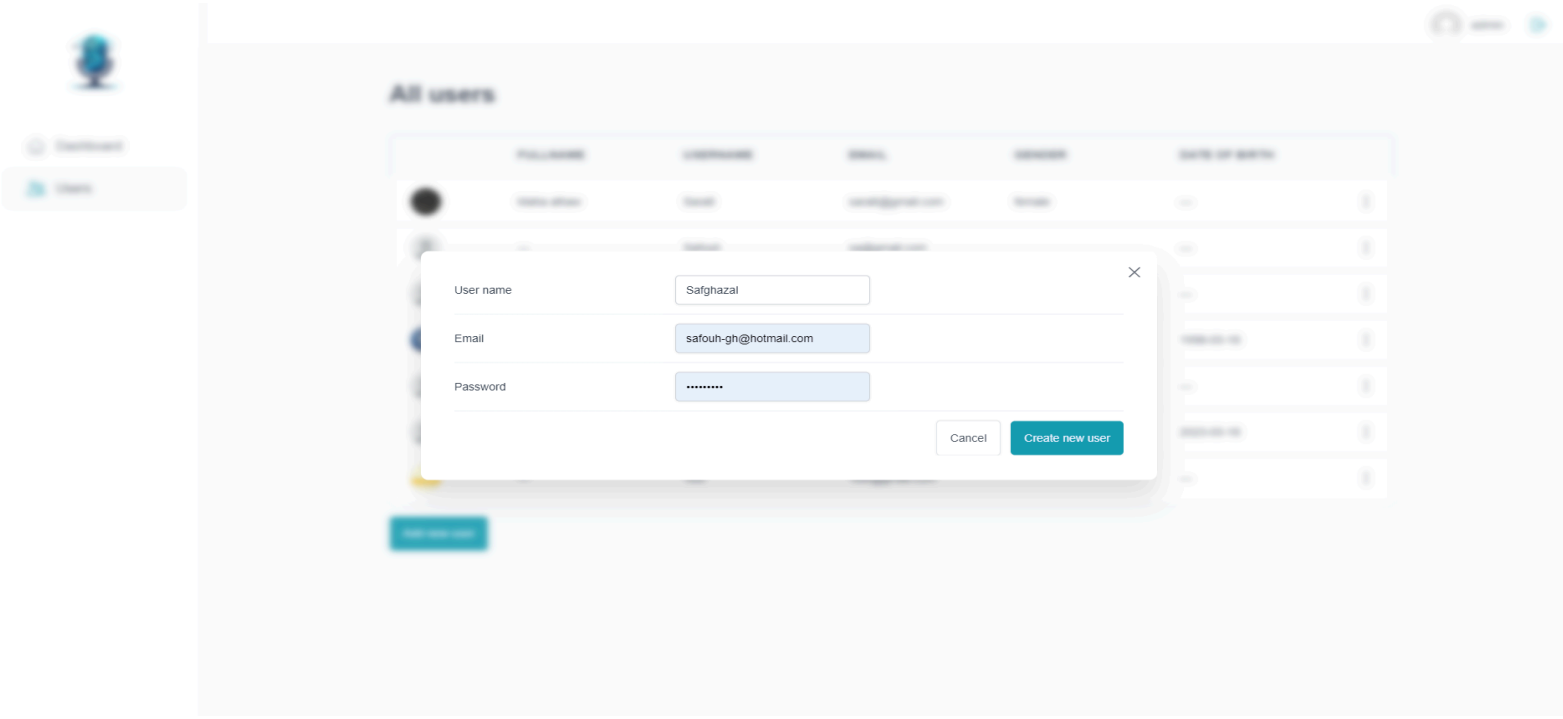


7.1.4. Users UI: *Figure [45-46]*

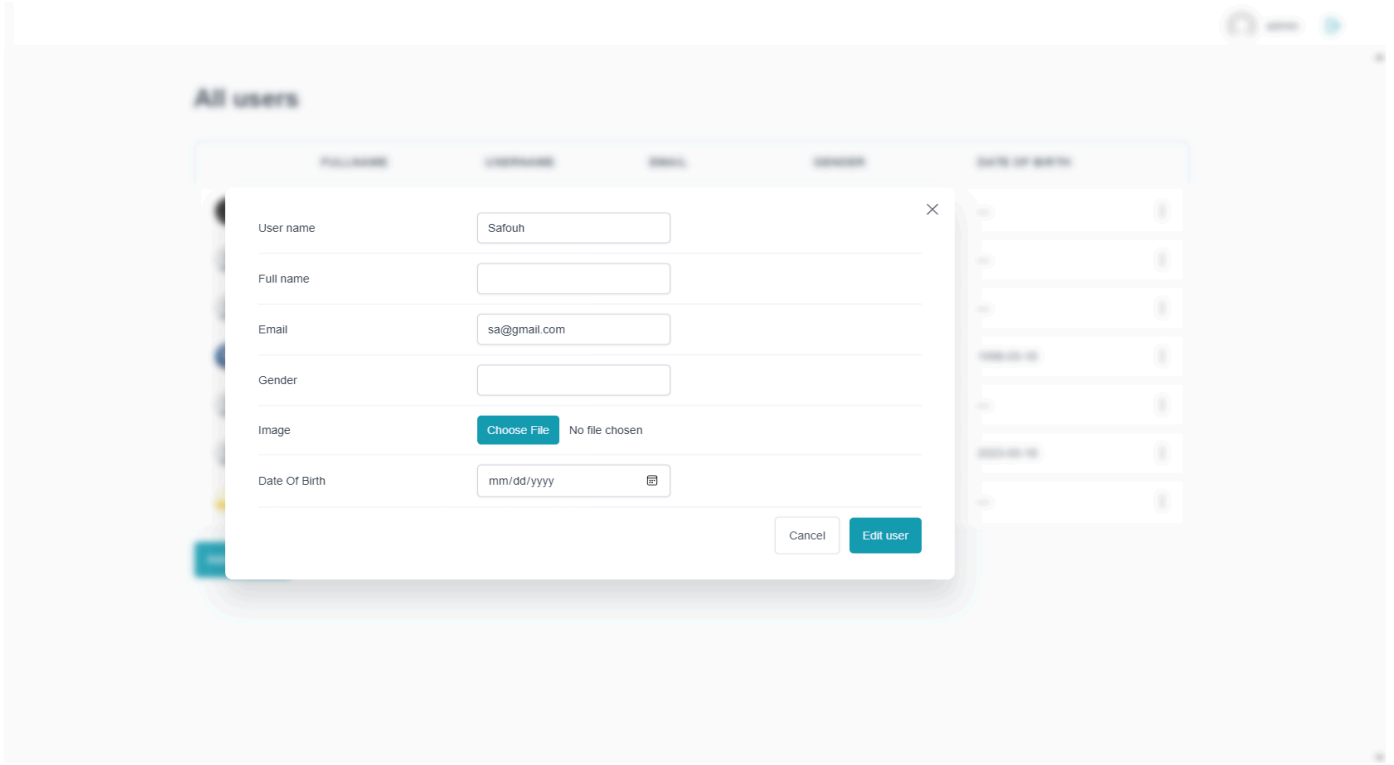




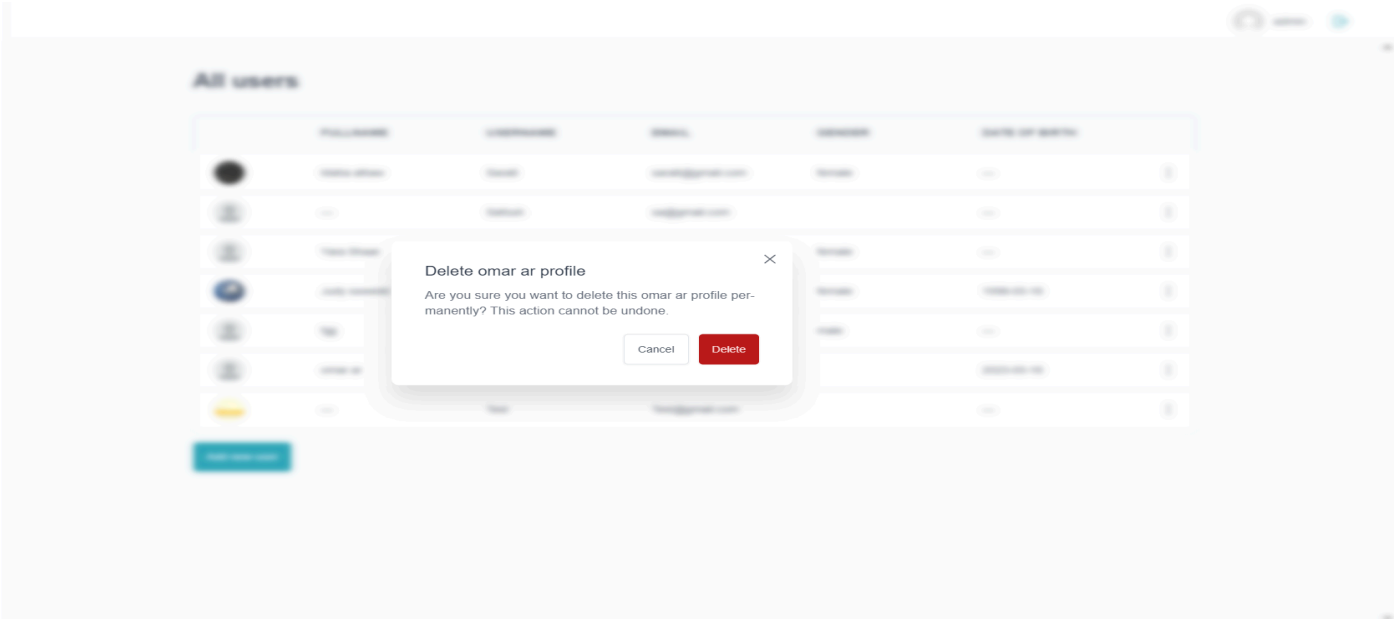
7.1.5 Add User: *Figure [47]*



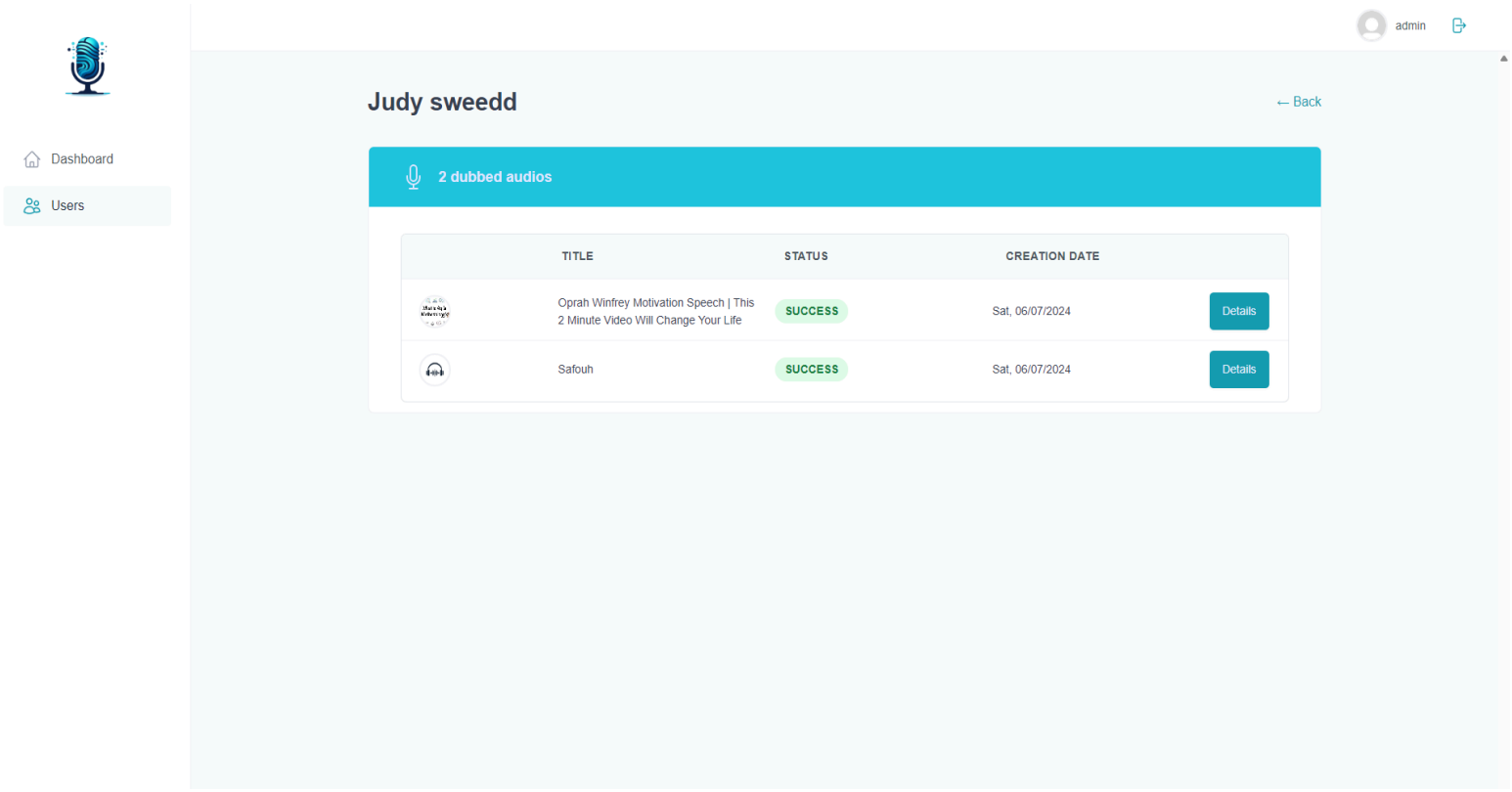
7.1.6 Update User: *Figure [48]*



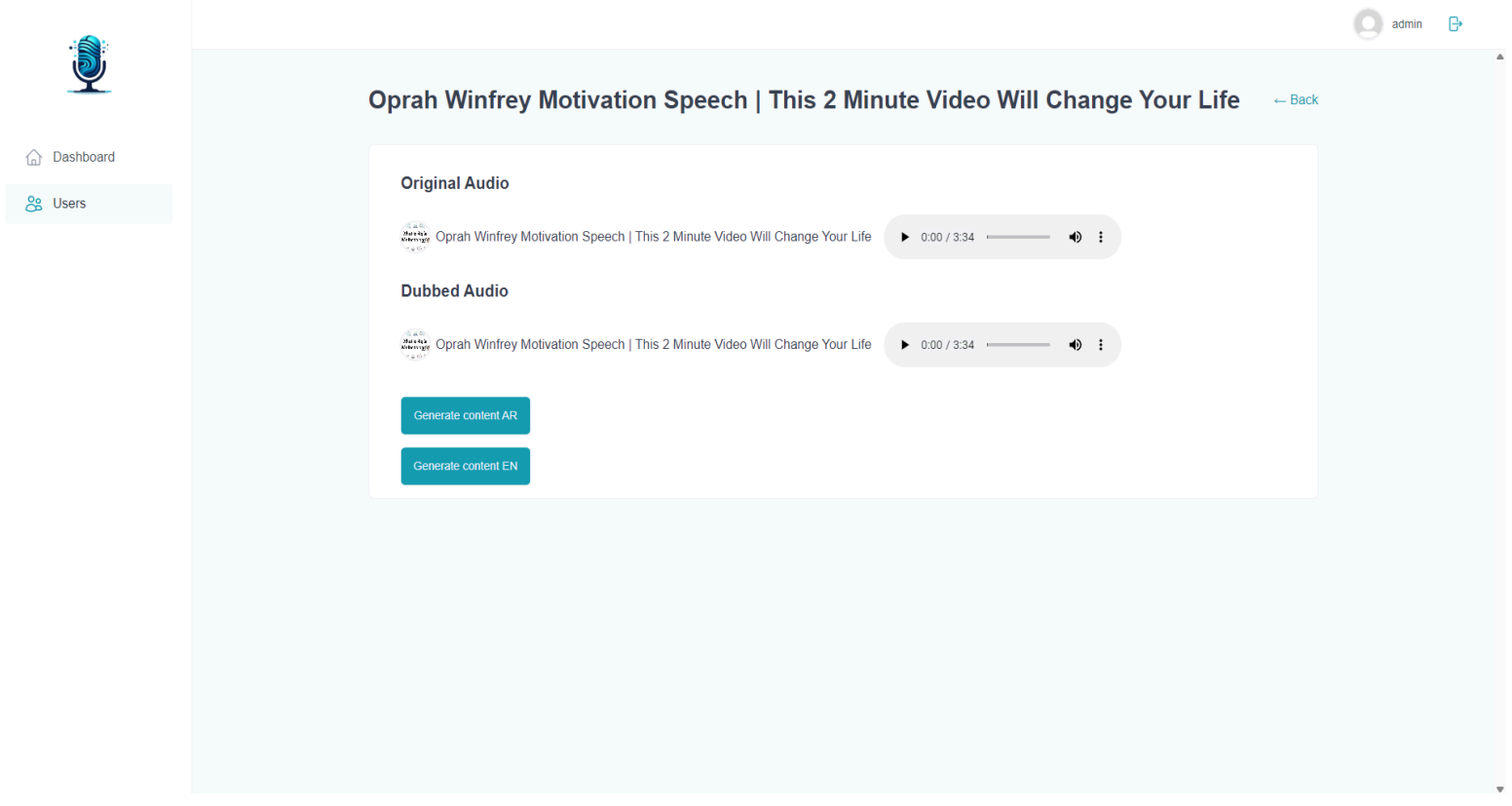
7.1.7 Delete User: *Figure [49]*



7.1.8 View History: *Figure [50]*

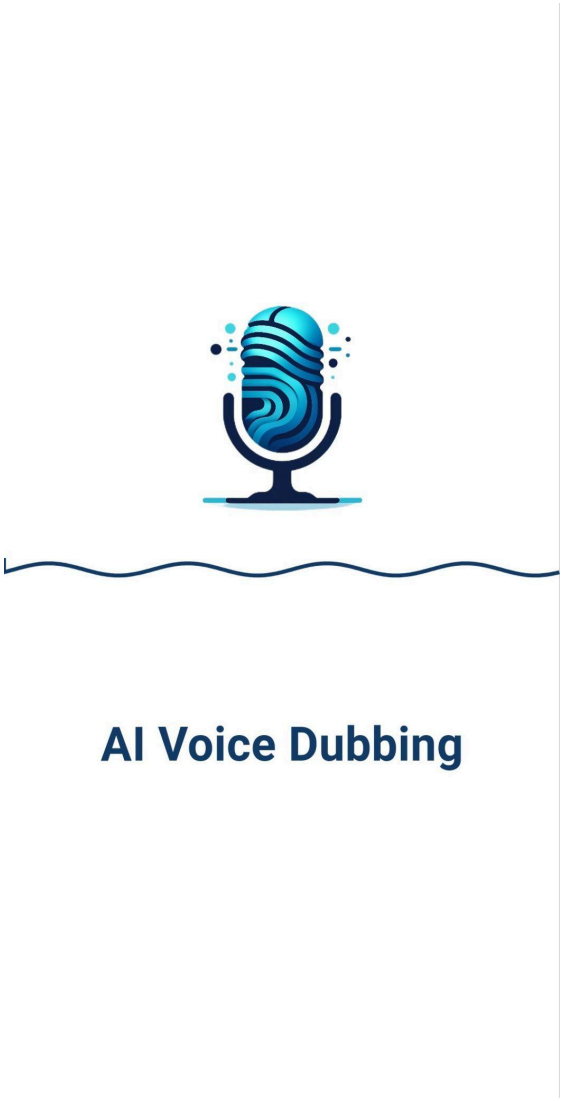


7.1.9 Details: *Figure [51]*

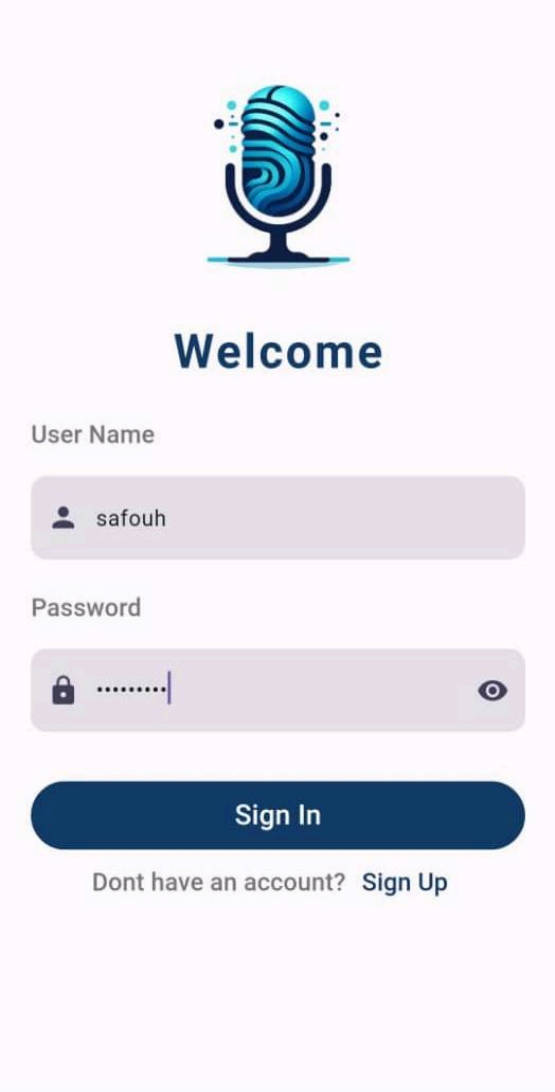


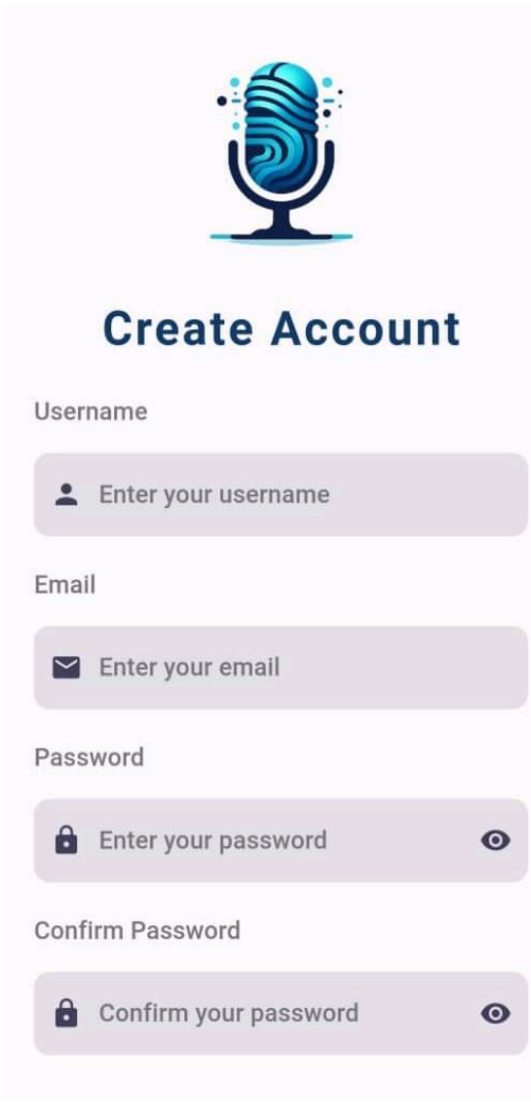
7.2. User Application

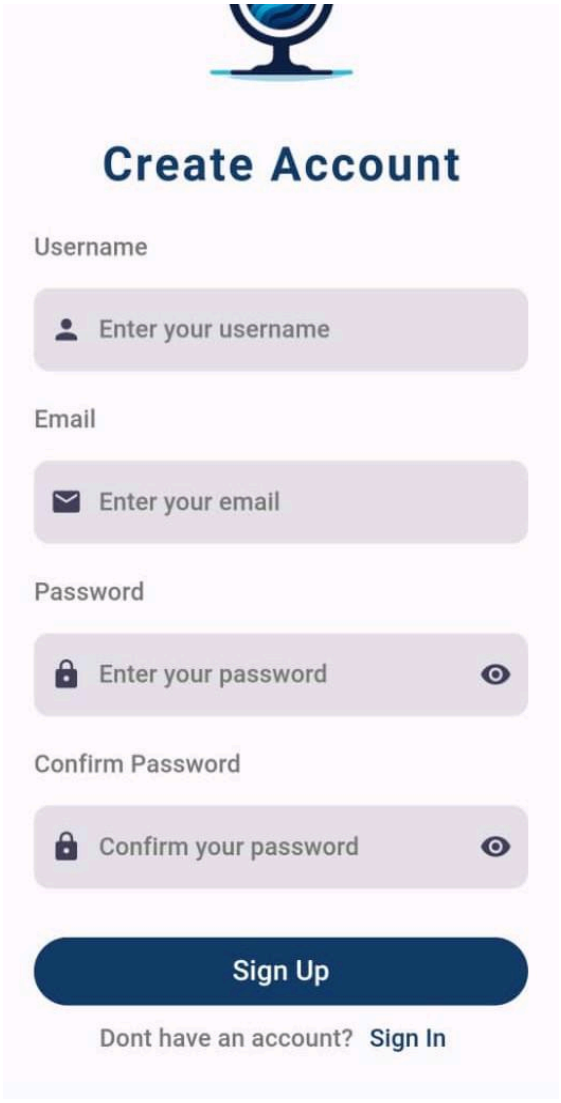
7.2.1. Splash Interface:



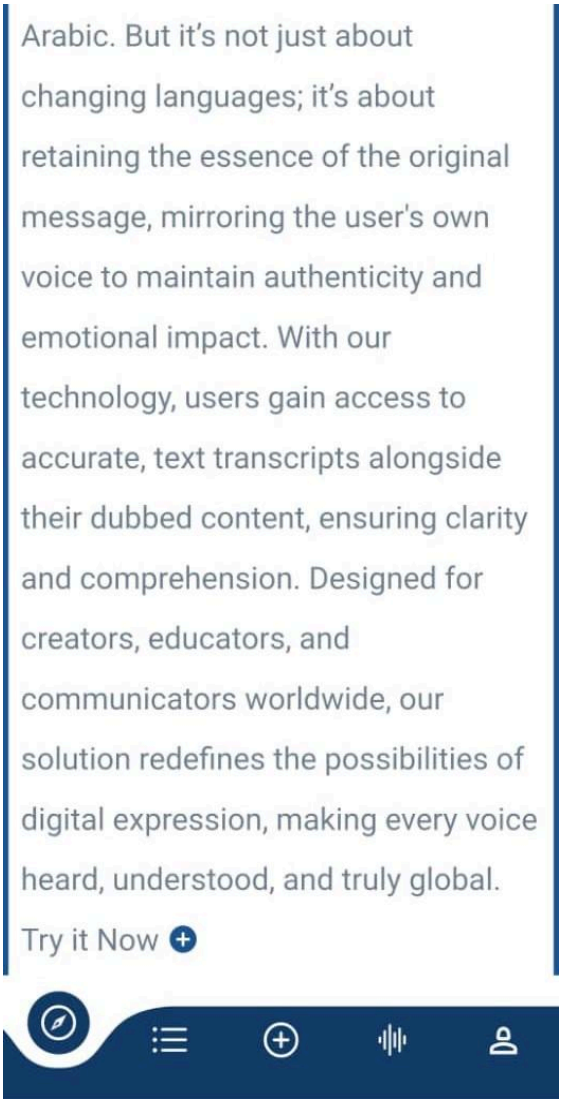
7.2.2. Login & Sign/up Interfaces:




Create Account
Username
Enter your username
Email
Enter your email
Password
Enter your password
Confirm Password
Confirm your password

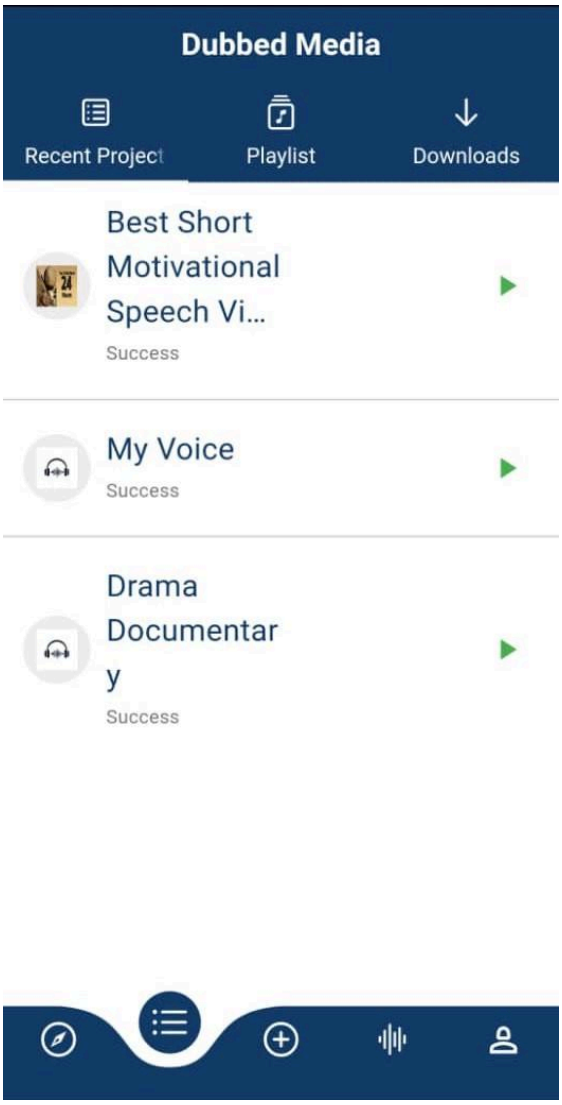

Create Account
Username
Enter your username
Email
Enter your email
Password
Enter your password
Confirm Password
Confirm your password
Sign Up
Dont have an account? Sign In

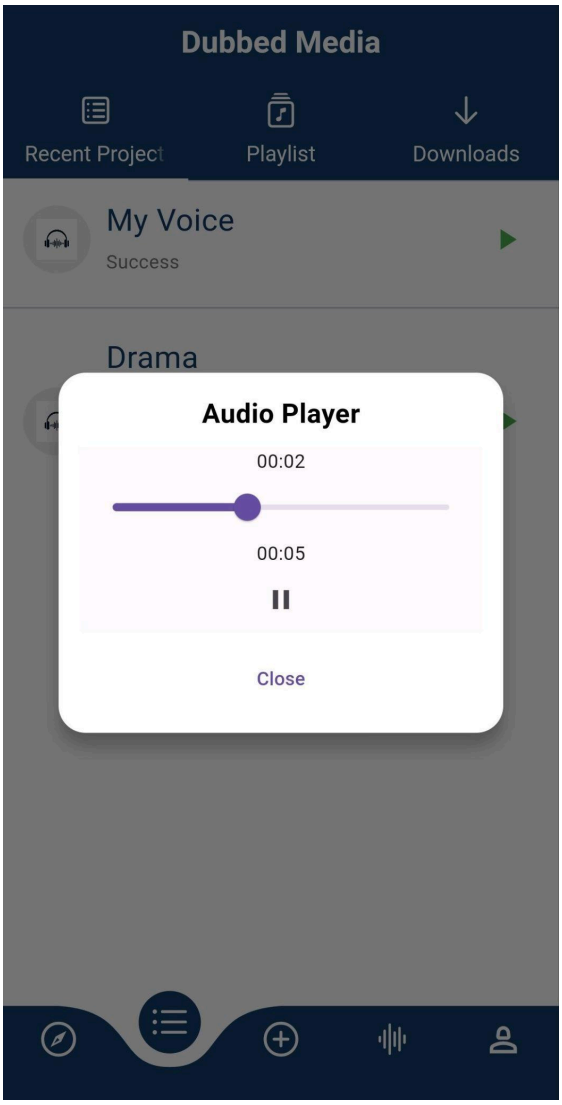
7.2.3. Explore Tab Interface:

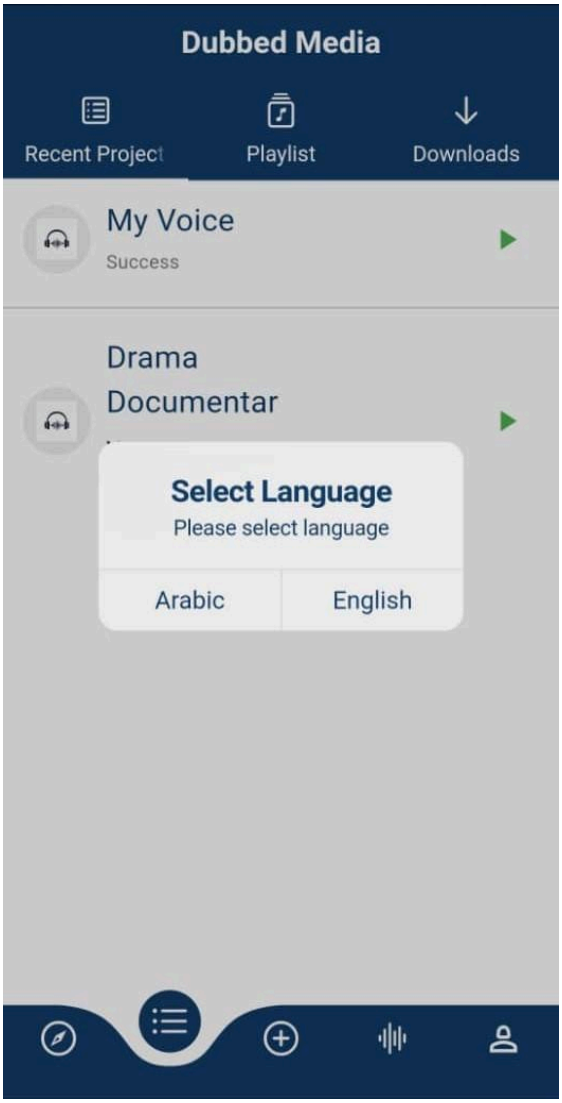


7.2.4. List Audios Tab

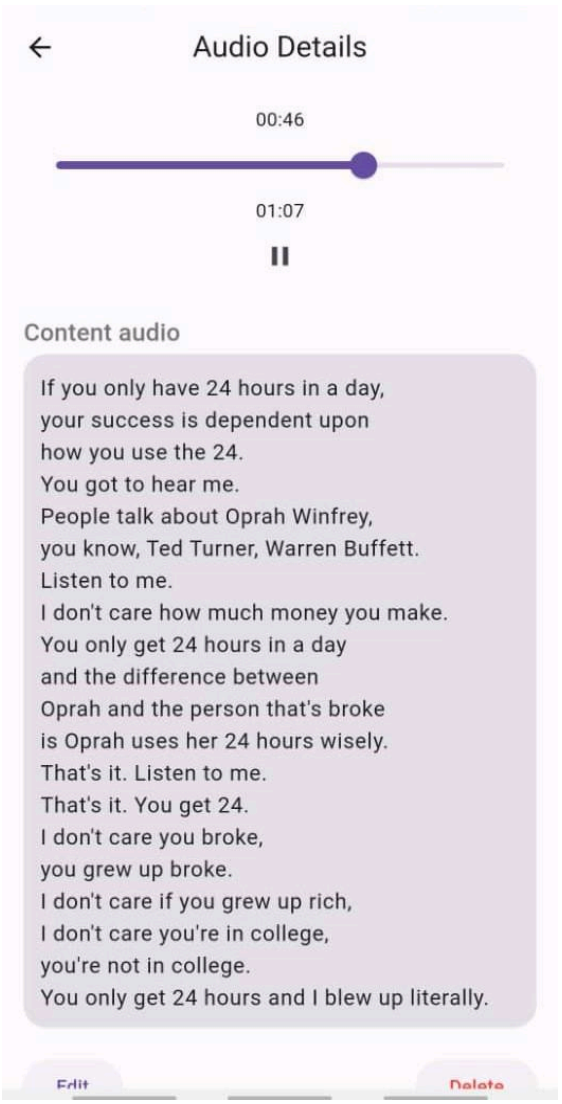
7.2.4.1. Recent Projects Interface:



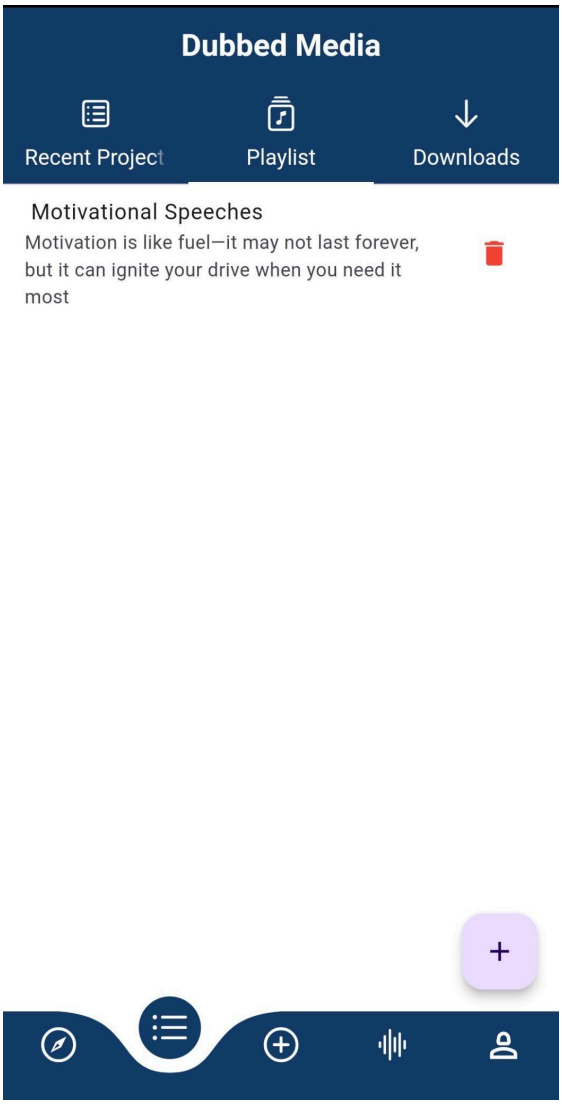


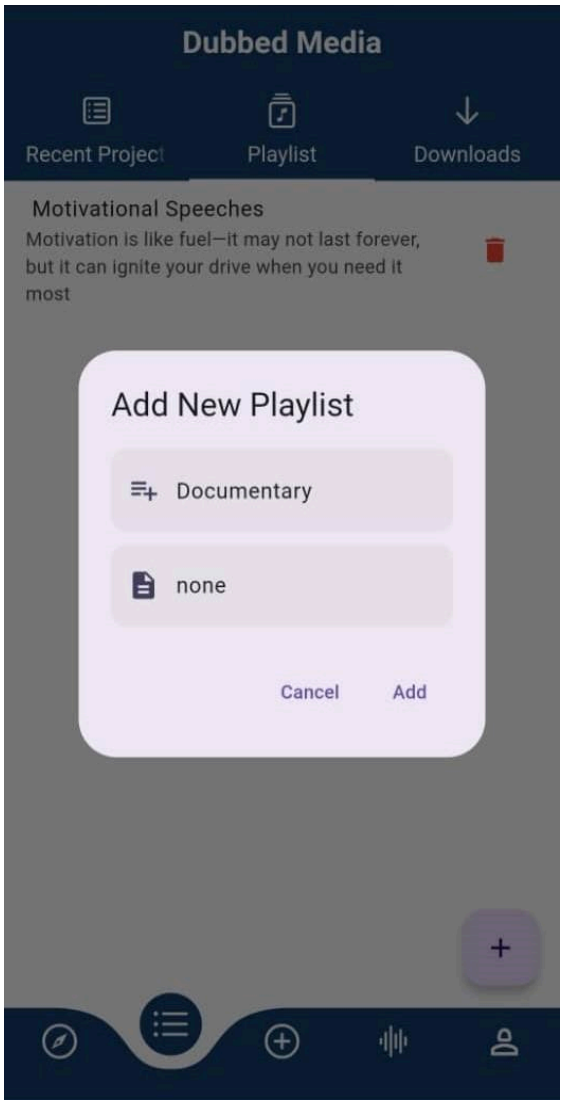

Dubbed Media
Recent Project
Playlist
Downloads
My Voice
Success
Drama
Documentar
Select Language
Please select language
Arabic
English

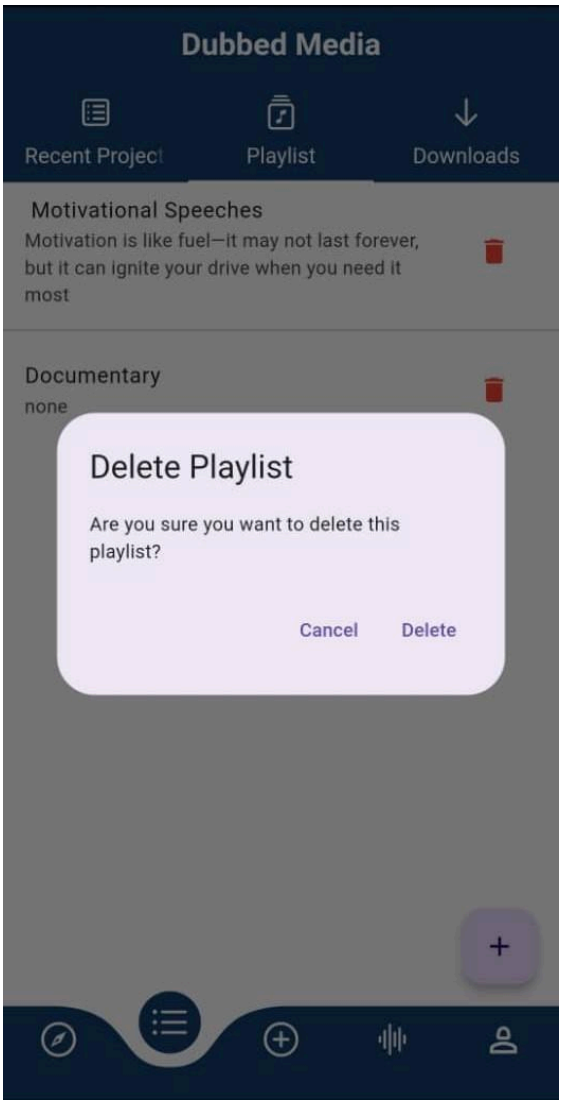

Audio Details
00:20
01:07
Content audio
إذا كان لديك فقط 24 ساعة في اليوم،
فإن نجاحك يعتمد على
كيفية استخدامك لتلك الـ 24 ساعة.
استمع إليّ.
الناس يتحدثون عن أوبرا وينفري،
تيد تيرنر، وورين بافيت.
استمع إليّ.
لا يهمني كم تربح من المال.
أنت تحصل على 24 ساعة في اليوم فقط
والفارق بين
أوبرا والشخص الفقير
هو أن أوبرا تستخدم 24 ساعة بحكمة.
هذا هو الأمر. استمع إليّ.
هذا هو الأمر. لديك 24 ساعة.
لا يهم إذا كنت مفلسًا،
أو نشأت في حالة فقر.
لا يهم إذا كنت نشأت في ثراء،
لا يهم إذا كنت في الجامعة،
أو ليس لديك جامعة.
لديك 24 ساعة فقط وأنا انفجرت حرفيًا.
Edit
Delete


Audio Details
00:46
01:07
Content audio
If you only have 24 hours in a day,
your success is dependent upon
how you use the 24.
You got to hear me.
People talk about Oprah Winfrey,
you know, Ted Turner, Warren Buffett.
Listen to me.
I don't care how much money you make.
You only get 24 hours in a day
and the difference between
Oprah and the person that's broke
is Oprah uses her 24 hours wisely.
That's it. Listen to me.
That's it. You get 24.
I don't care you broke,
you grew up broke.
I don't care if you grew up rich,
I don't care you're in college,
you're not in college.
You only get 24 hours and I blew up literally.

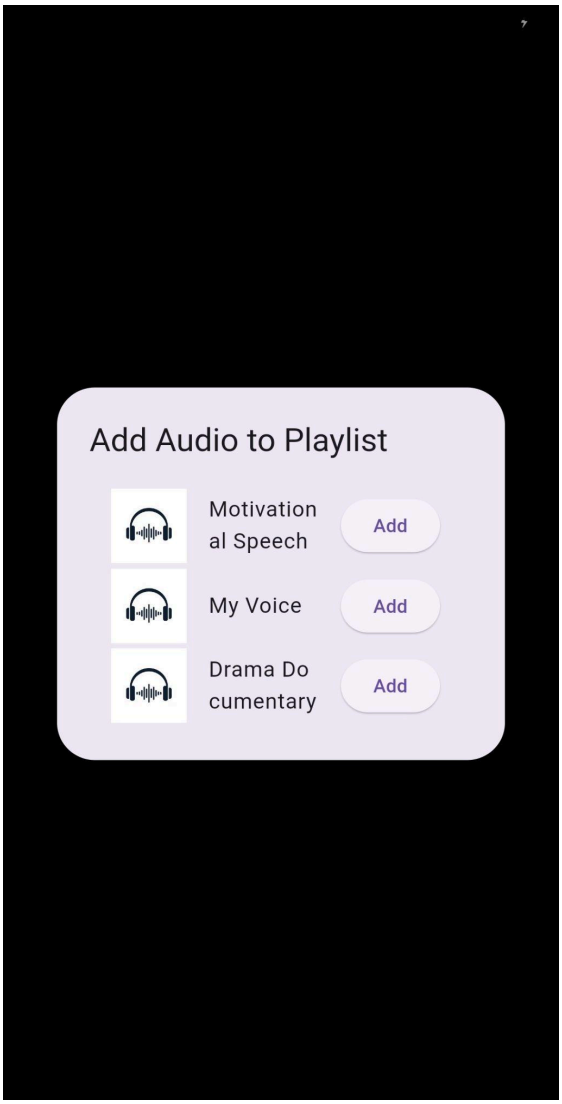
7.2.4.2. Playlists Interface:



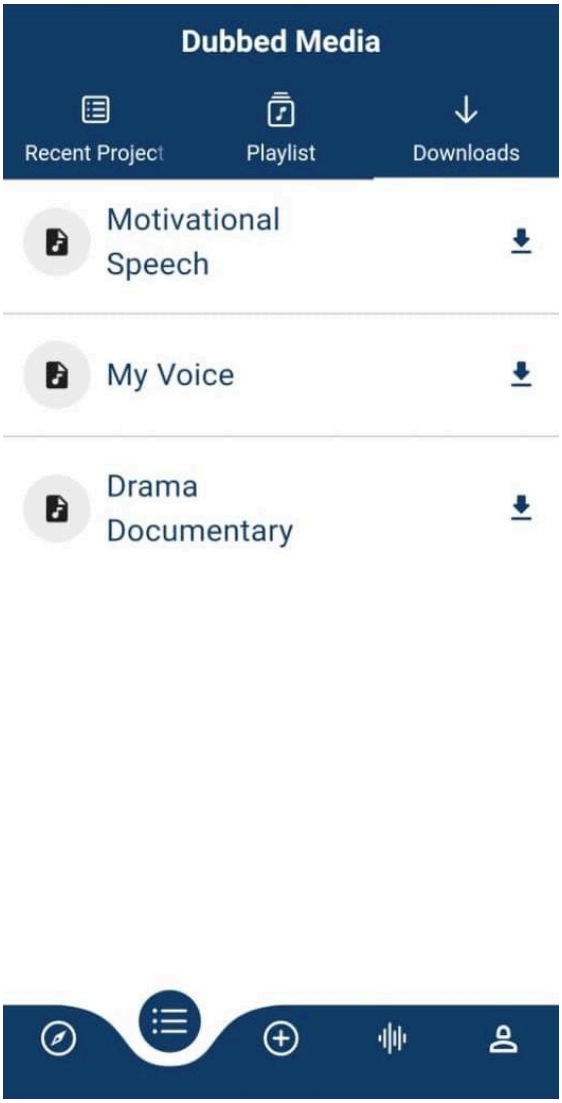



Dubbed Media
Recent Project
Playlist
Downloads
Motivational Speeches
Motivation is like fuel—it may not last forever, but it can ignite your drive when you need it most
Documentary
none
Delete Playlist
Are you sure you want to delete this playlist?
Cancel
Delete

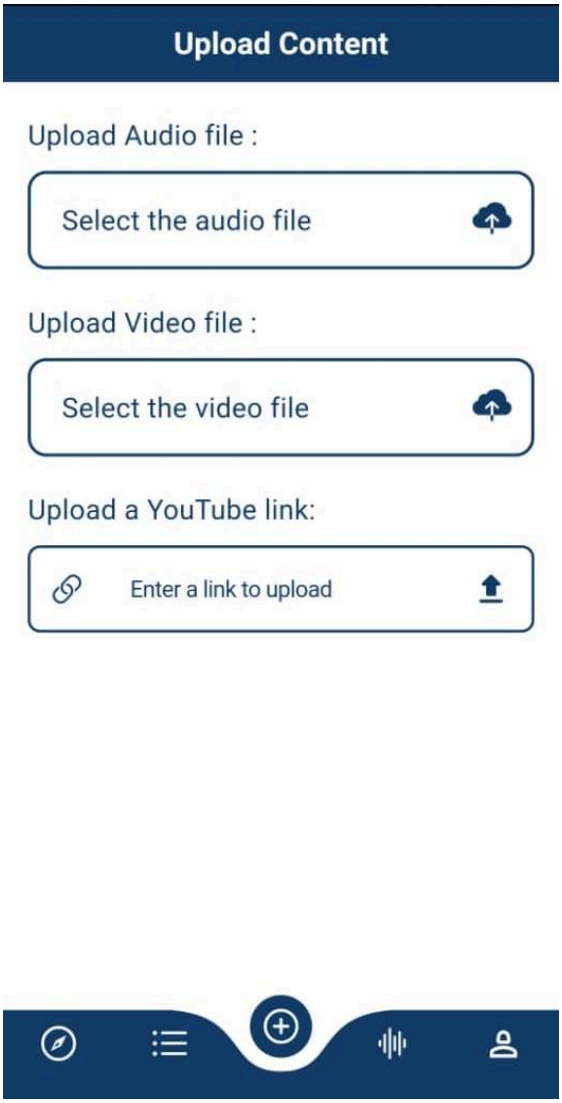

Playlist
Best Short Motivational Spe...

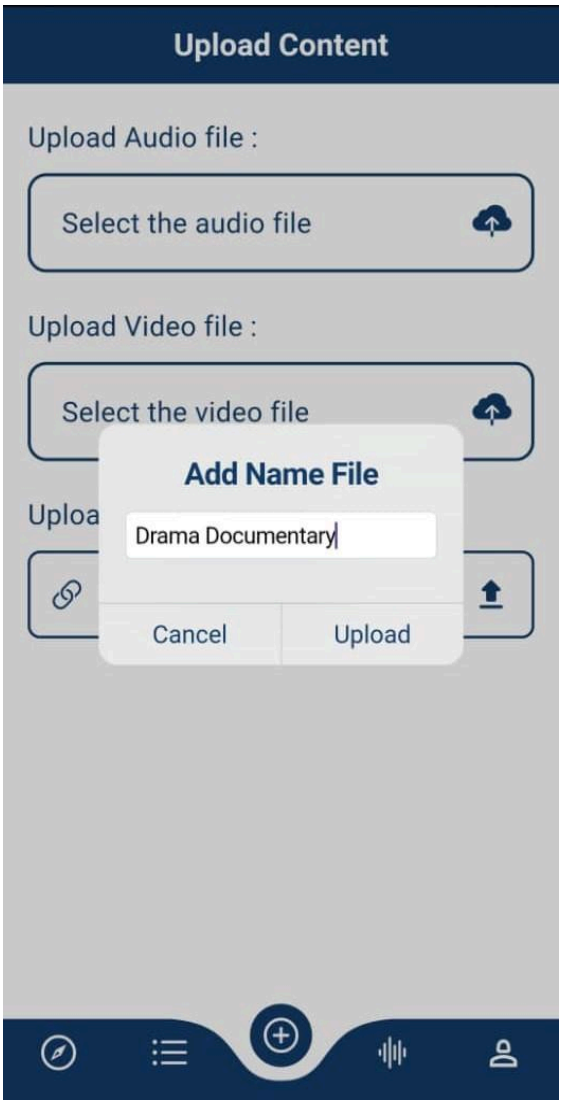

Add Audio to Playlist
Motivation al Speech
Add
My Voice
Add
Drama Do cumentary
Add

7.2.4.3. Downloads Interface(with downloaded content):

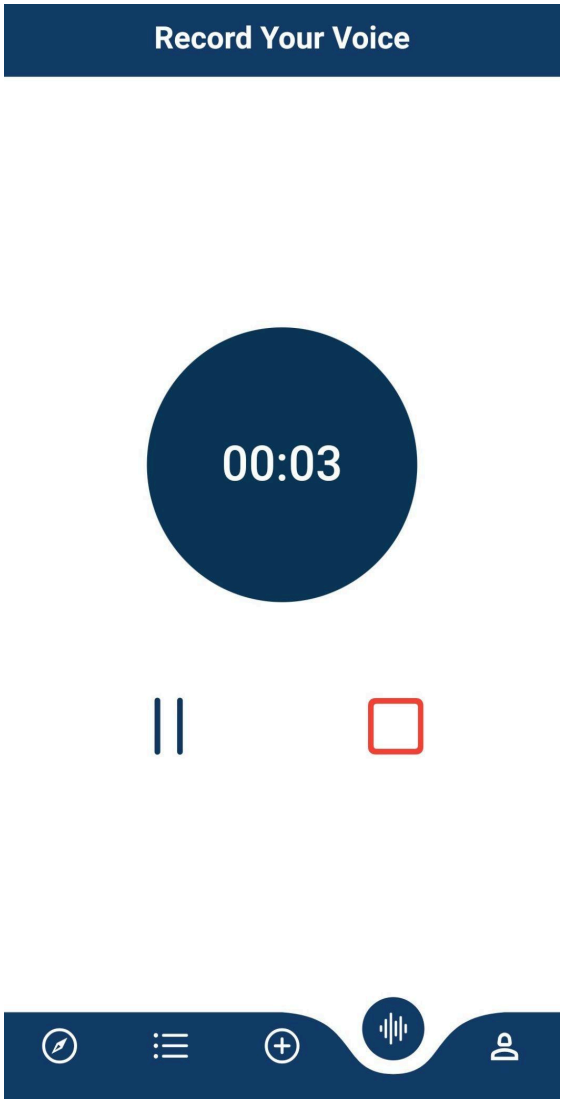


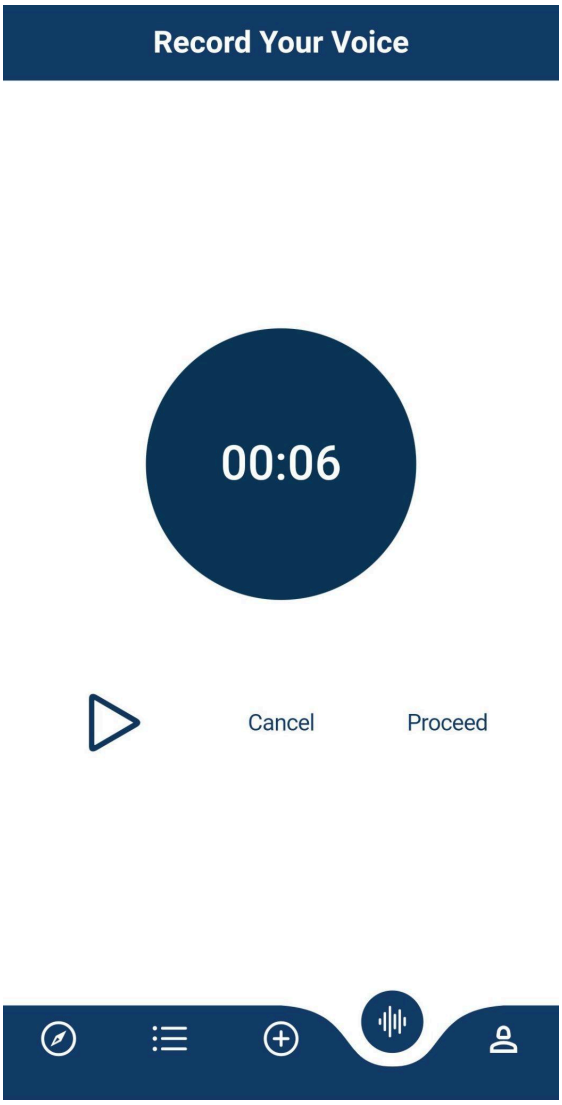
7.2.5. Upload Media Interface:

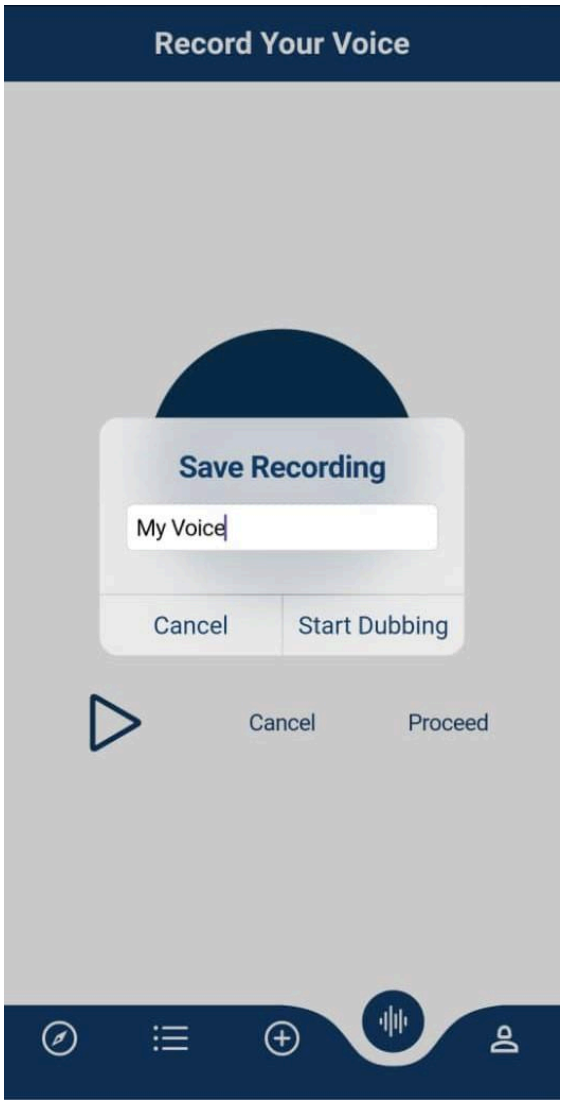




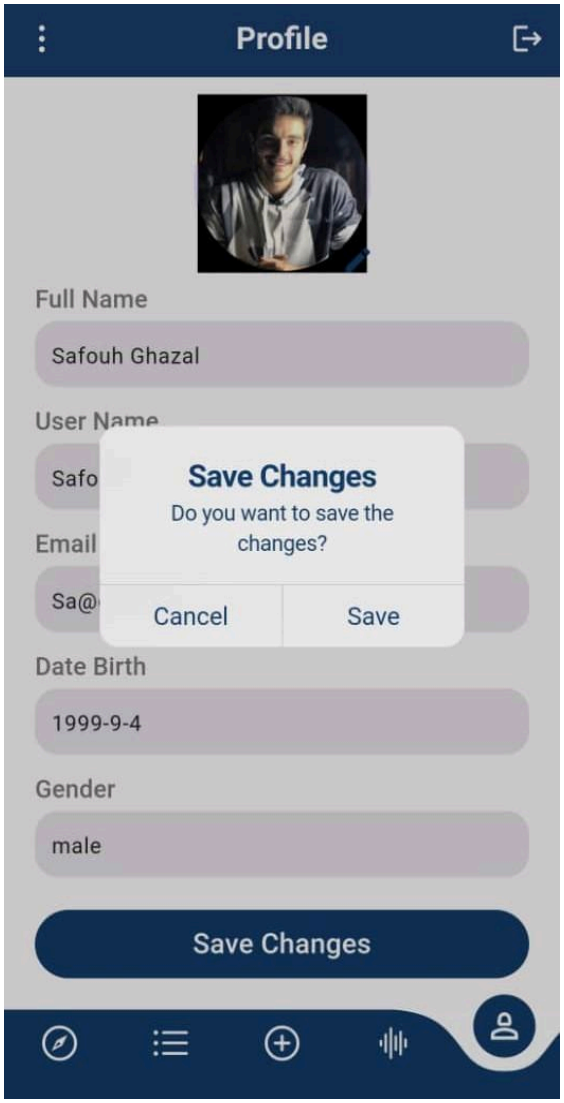
7.2.6. Voice Recorder Interface:

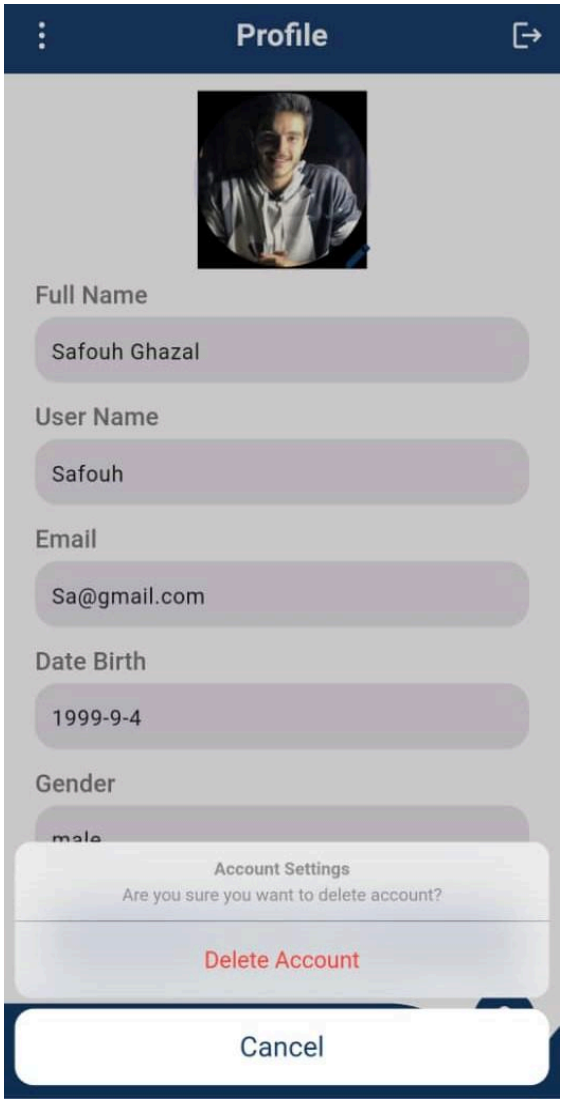



Record Your Voice
00:06
Cancel
Proceed


Record Your Voice
Save Recording
My Voice
Cancel
Start Dubbing
Cancel
Proceed

7.2.7. View Profile Interface:


Profile
Full Name
Safouh Ghazal
User Name
Safo
Email
Sa@
Save Changes
Do you want to save the changes?
Cancel
Save
Date Birth
1999-9-4
Gender
male
Save Changes


Profile
Full Name
Safouh Ghazal
User Name
Safouh
Email
Sa@gmail.com
Date Birth
1999-9-4
Gender
Account Settings
Are you sure you want to delete account?
Delete Account
Cancel

Chapter 8: Testing & Discussion

8.1 English Speech-to-Text

8.1.1 STT From Scratch

In my studies, I conducted numerous experiments to achieve suitable results. The models were optimized using the Adam optimizer. For the Convolutional Neural Network (CNN), the ReLU activation function was employed, while the fully connected layer utilized the softmax function. The performance was measured using Word Error Rate (WER) and Character Error Rate (CER). The Connectionist Temporal Classification (CTC) loss function was used due to the sequential nature of the problem, which resulted in the high loss values mentioned in the table below.

Model	Training Loss	Loss
5	69.7014	86.32
8	19.3358	88.3

8.1.2 STT Pretrained

In this study, several pre-trained models, including Whisper and Wave2Vec, were utilized,each employing custom optimizers, activation functions, and loss functions tailored to their specific tasks. Whisper and Wave2Vec were optimized using the Adam optimizer. Whisper demonstrated top performance due to its lower validation loss and superior metrics, including the Weighted Error Rate (WER), compared to Wave2Vec. Wave2Vec, optimized similarly but employing a connectionist temporal classification (CTC) loss function suitable for its speech-to-text tasks, underperformed with a validation loss of 0.84, indicating underfitting issues. The table below encapsulates the outcomes

from each model implemented in this investigation, highlighting Whisper's efficacy in delivering accurate speech-to-text conversions:

Model	WER	CER	Loss
Whisper	8.835079	4.539610	0.005200
wave2vec	1.0	1.0	0.840459

Results of English Speech to text

8.2 Text-to-Speech and Voice Cloning

8.2.1 Testing the Model

The **MOS** is calculated as the arithmetic mean over single ratings performed by human subjects for a given stimulus in a subjective quality evaluation test. Thus:

Rating	Label
5	Excellent
4	Good
3	Fair
2	Poor

$$MOS = \frac{\sum_{n=1}^{N} R_n}{N} \; ; N = 10$$

1	Bad

Where *R* are the individual ratings for a given stimulus by *N* subjects.

Experiment 1: The first experiment was mainly for testing the

XTTS model capabilities

- The audio used was a sample the model offered.
- The text was a very short Text made of only 5 words.
- No preprocessing was done.

MOS = 4.5

Experiment 2: The second experiment was done to test the model performance to unknown audio

- The audio used was a clear speaker audio with no background sounds
- The text was a very short text made of only 5 words.
- No preprocessing was done.
- No post processing was done.

MOS = 4.3 Wrong spelling for some of the non diacritic words

Experiment 3: The third experiment was done to test the model ability to read diacritics

- The audio used was a clear speaker audio with no background sounds
- The text was a very short text made of only 5 words.
- Diacritization on words were added manually
- No post processing was done.

MOS = 4.5

Experiment 4: The fourth experiment was done to test the model's text input limits

- The audio used was a sample the model offered.
- The text was 700 words long.
- No preprocessing was done.
- No post processing was done.

The model failed to give an output it only accepts a 400 character (spaces included)

Experiment 5: The fifth experiment was done to solve the previous experiment's error

- The audio used was a sample the model offered.
- The text was 700 words long.
- Only sentence splitting preprocessing.
- Audio list concatenation post processing.

MOS = 4 The sentence was cut mid-sentence when spoken.

Experiment 6: The sixth experiment was done to solve the previous experiment's error

- The audio used was a sample the model offered.
- The text was 700 words long.
- Only sentence tokenization and splitting preprocessing.
- Audio list concatenation post processing.

MOS = 4.3

Experiment 7: The seventh experiment was done to test the model's audio limits

- The audio used was an audio with background sounds and music.
- The text was 700 words long.
- Only sentence tokenization and splitting preprocessing.
- Audio list concatenation post processing.

The output given was unclear and filled with noise (no spoken words were able to be heard)

Experiment 8: The eighth experiment was done to solve the previous experiment's error

- The audio used was an audio with background sounds and music.
- The text was 700 words long.
- Only sentence tokenization and splitting preprocessing. Audio preprocessing was done using only librosa's functions
- Audio list concatenation post processing.

MOS = 2.5 Some words were able to be spoken by the model and the speaker voice was not close the original speaker's voice

Experiment 9: The ninth experiment was done to solve the previous experiment's error

- The audio used was an audio with background sounds and music.
- The text was 700 words long.
- Sentence tokenization and splitting preprocessing. Audio preprocessing was done using a pre-trained audio splitting tool.
- Audio list concatenation post processing.

MOS = 3.8

Experiment 10 : The tenth experiment was the same as the previous experiment but diacritics were added.

- The audio used was an audio with background sounds and music.
- The text was 700 words long.
- All preprocessing steps mentioned before were implemented.
- Audio list concatenation post processing.

MOS = 3.9

Experiment 11 : The eleventh experiment was done using a music audio

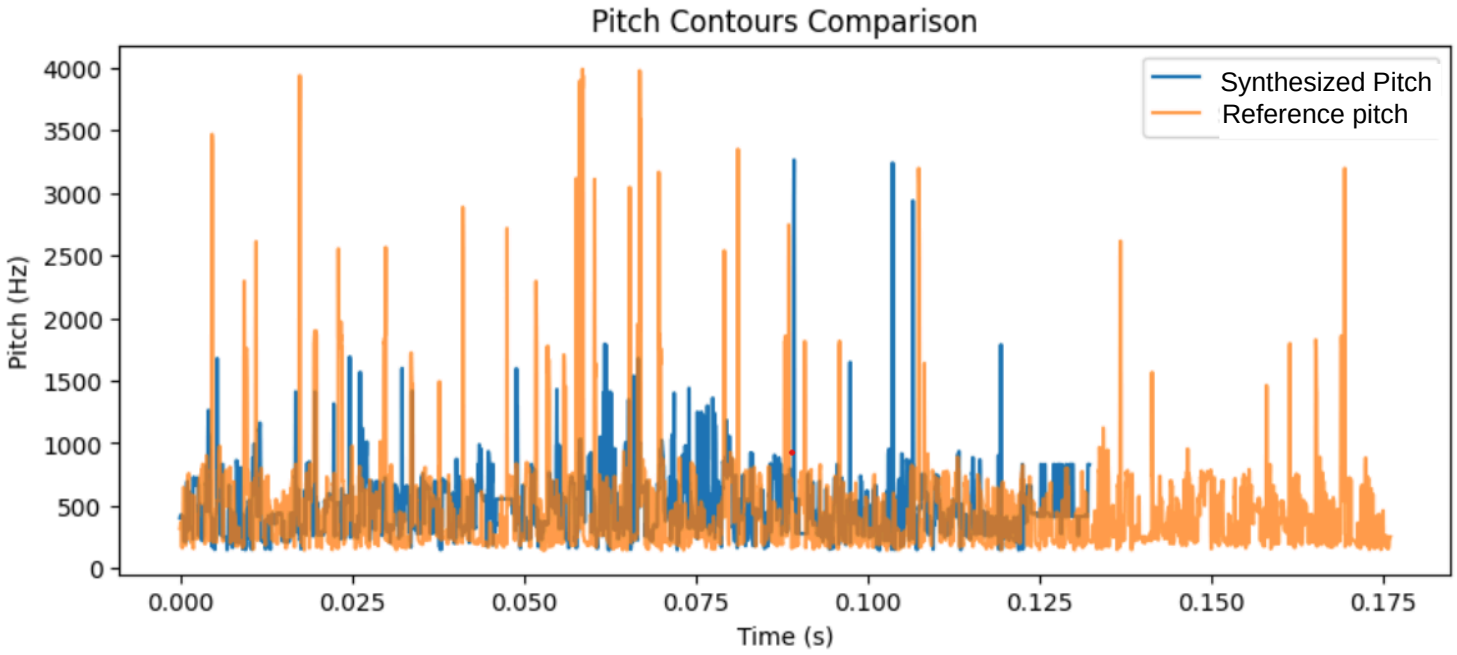
- The audio used was an audio with background sounds and music.
- The text was 1392 words long.
- All preprocessing steps mentioned before were implemented.
- Audio list concatenation post processing.

MOS = 3 The output speaker's voice was not very close to the original speaker's

Experiment 12 : The twelfth experiment was done using a motivational speaker's audio

- The audio used was an audio with background sounds and music.
- The text was 2173 words long.
- All preprocessing steps mentioned before were implemented.
- Audio list concatenation post processing.

MOS = 4.1



Pitch Comparison between the speaker's original voice and the output speaker's voice

For the final experiment

8.2.1 Discussion

Our evaluation process revealed several areas for potential improvement in the model's capabilities:

- **Limited Input Length:** The current model can only process sentences with a maximum of 400 characters (including spaces). This restriction might be inconvenient for handling longer sentences.
- **Sensitivity to Diacritics:** The model's performance might be impacted by the presence or absence of diacritics (vowel markings) in the input text. This could lead to spelling errors in synthesized speech if the text lacks diacritics.
- **Background Noise Interference:** The model may struggle to accurately capture and reproduce speaker voice characteristics if the original audio contains background noise.

Conclusion and Future Works

The advancements in AI-driven voice dubbing technologies represent a significant stride towards dismantling language barriers in multimedia content. These innovations not only enhance the accessibility of information across different linguistic demographics but also preserve the emotional and cultural nuances often lost in traditional translation methods. By utilizing advanced digital algorithms, these systems can generate dubbed audio that is not only linguistically accurate but also tonally and contextually appropriate. Additionally, they incorporate voice cloning techniques to maintain the original speaker's voice, ensuring authenticity. This capability is crucial for providing non-native speakers with an immersive and comprehensible viewing experience, making content universally enjoyable and accessible.

Looking ahead, the platform focuses on enhancing dubbing technology and community involvement, promising opportunities for advancements in voice synthesis and performance.

Possible features to be added later on:

1. **Multilingual Dubbing:** Expanding capabilities beyond Arabic to support dubbing in various languages.
2. **Emotional Text-to-Speech:** Enabling synthesized speech to convey emotions through incorporating emotional datasets into training models.
3. **Real-Time Dubbing:** This would allow viewers to enjoy live broadcasts and streams in their native language with minimal delay. Achieving this requires significant progress in real-time speech recognition, translation, voice cloning, and low-latency audio delivery.
4. **Video Dubbing:** Synchronizing synthesized speech with video content, potentially including lip movement manipulation for a natural effect.

References

[0] Casanova, Edresson, et al. "XTTS: a Massively Multilingual Zero-Shot Text-to-Speech Model." arXiv preprint arXiv:2406.04904 (2024).

[1] J. Betker, “Better speech synthesis through scaling,” arXiv preprint arXiv:2305.07243, 2023.

[2] P. Gage, “A new algorithm for data compression,” C Users Journal, vol. 12, no. 2, pp. 23–38, 1994.

[3] J.-B. Alayrac, J. Donahue, P. Luc, A. Miech, I. Barr, Y. Hasson, K. Lenc, A. Mensch, K. Millican, M. Reynolds et al., “Flamingo: a visual language model for few-shot learning,” Advances in Neural Information Processing Systems, vol. 35, pp. 23 716–23 736, 2022.

[4] E. Casanova, J. Weber, C. D. Shulby, A. C. Junior, E. Golge, and ¨ M. A. Ponti, “Yourtts: Towards zero-shot multi-speaker tts and zero-shot voice conversion for everyone,” in International Conference on Machine Learning. PMLR, 2022, pp. 2709–2720.

[5] J. Kong, J. Kim, and J. Bae, “Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis,” arXiv preprint arXiv:2010.05646, 2020.

[6] H. S. Heo, B.-J. Lee, J. Huh, and J. S. Chung, “Clova baseline system for the voxceleb speaker recognition challenge 2020,” arXiv preprint arXiv:2009.14153, 2020.

[7] M. A. Ahmed and S. Trausan-Matu, "Using natural language processing for

analyzing Arabic poetry rhythm," 2017 16th RoEduNet Conference: Networking in Education and Research (RoEduNet), Targu-Mures, Romania, 2017, pp. 1-5, doi: 10.1109/ROEDUNET.2017.8123759.

[8] AlShammari, Norah, and Amal AlMansour. "Aspect-based sentiment analysis and location detection for Arabic language Tweets." Applied Computer Systems 27.2 (2022): 119-127.

[9] Wikipedia, 2013,” Speech_translation”. https://en.m.wikipedia.org/wiki/Speech_translation

[10] Sandeep Dhawan. (2022),” Speech To Speech Translation: Challenges and Future”. https://ijcat.com/archieve/volume11/volume11issue3.pdf

[11] ”voice recognition (speaker recognition)”, (2019). https://www.techtarget.com/searchcustomerexperience/definition/voice-recognition-speaker-recognition

[12] ”Machine Translation”,(2024). https://www.studysmarter.co.uk/explanations/english/linguistic-terms/machine-translation

[13] Wikipedia, 2024,” Speech synthesis”. https://en.wikipedia.org/wiki/Speech_synthesis

[14] Kim Martin, VP Marketing(2023). ” What is Voice Cloning?”. https://www.idrnd.ai/what-is-voice-cloning/

[15] ”Voice Cloning: What It Is & How to Get Your VoiceCloned”. 2024 https://podcastle.ai/blog/what-is-voice-cloning

[16] ”Everything you need to know about voice cloning”. 2024
https://deepgram.com/learn/voice-cloning-everything-to-know

[17] Mirco Ravanelli , Titouan Parcollet, Peter Plantinga , Aku Rouhe , Samuele Cornell, Loren Lugosch , Cem Subakan , Nauman Daulatabad , Abdelwahab Heba , Jianyuan Zhong , Ju-Chieh Chou , Sung-Lin Yeh , Szu-Wei Fu , Chien-Feng Liao , Elena Rastorgueva , François Grondin , William Aris, Hwidong Na , Yan Gao , Renato De Mori, & Yoshua Bengio
,(2021,8 june). SpeechBrain: A General-Purpose Speech Toolkit
https://arxiv.org/pdf/2106.04624

[18] Awni Hannun , Carl Case, Jared Casper, Bryan Catanzaro, Greg Diamos, Erich Elsen, Ryan Prenger, Sanjeev Satheesh, Shubho Sengupta, Adam Coates, Andrew Y. Ng. (2014,Dec 19).Deep Speech: Scaling up end-to-end speech recognition.
https://arxiv.org/pdf/1412.5567v2

[19] F. Matern, C. Riess & M. Stamminger(2012, apr 6), SPGISpeech: 5,000 hours of transcribed financial audio for fully formatted end-to-end speech recognition.
https://arxiv.org/pdf/2104.02014v2

[20] Yajie Miao, Mohammad Gowayyed & Florian Metze . (2015,june 18), EESEN: END-TO-END SPEECH RECOGNITION USING DEEP RNN MODELS AND WFST-BASED DECODING https://arxiv.org/pdf/1507.08240v3

[21] Hamad, M., & Hussain, M. (2011). Arabic Text-To-Speech Synthesizer. (2011) IEEE Student Conference on Research and Development. https://sci-hub.ru/10.1109/SCOReD.2011.6148774

[22] Amr Youssef , Ossama Emam. (2004,april 21). An Arabic TTS System Based on the IBM Trainable Speech Synthesizer. https://www.afcp-parole.org/doc/Archives_JEP/2004_XXVe_JEP_Fes/actes/arabe2004/PAAY01.pdf

[23] Hawau Olamide Toyin ,Amirbek Djanibekov ,Ajinkya Kulkarni, Hanan Aldarmaki.(2023). ArTST: Arabic Text and Speech Transformer. https://arxiv.org/pdf/2310.16621

[24] Yixuan Zhou, Changhe Song, Xiang Li, Luwen Zhang, Zhiyong Wu, Yanyao Bian, Dan Su, Helen Meng. (11,Nov 2022). Content-Dependent Fine-Grained Speaker Embedding for Zero-Shot Speaker Adaptation in Text-to-Speech Synthesis. https://arxiv.org/pdf/2204.00990

[25] Joun Yeop Lee, Myeonghun Jeong, Minchan Kim, Ji-Hyun Lee, Hoon-Young Cho, Nam Soo Kim (25 June 2024). High Fidelity Text-to-Speech Via Discrete Tokens Using Token Transducer and Group Masked Language Model. https://arxiv.org/pdf/2406.17310

[26] Edresson Casanova1, Kelly Davis, Eren G¨olge, G¨orkem G¨oknar, Iulian Gulea, Logan Hart, Aya Aljafari1, Joshua Meyer, Reuben Morais, Samuel Olayemi, and Julian Weber (7 Jun 2024) XTTS: a Massively Multilingual Zero-Shot Text-to-Speech Model

[27] He, Ren Yu. "Design and implementation of web based on Laravel framework." 2014 International Conference on Computer Science and Electronic Technology (ICCSET 2014). Atlantis Press, 2015.

[28] Westerveld, Dave. API Testing and Development with Postman: A practical guide to creating, testing, and managing APIs for automated software testing. Packt Publishing Ltd, 2021.

[29] Gackenheimer, Cory. Introduction to React. Apress, 2015.

[30] Wu, Wenhao. "React Native vs Flutter, Cross-platforms mobile application frameworks." (2018).

[31] Christudas, Binildas, and Binildas Christudas. MySQL. Apress, 2019.

[32] Bisong, Ekaba, and Ekaba Bisong. "Google colaboratory." Building machine

learning and deep learning models on google cloud platform: a comprehensive guide for beginners (2019): 59-64.

[33] Islam, Quazi Nafiul. Mastering PyCharm. Packt Publishing Ltd, 2015.

[34] Zengyi Qin Wenliang Zhao Xumin Yu Xin Sun (2 Jan 2024) OpenVoice: Versatile Instant Voice Cloning https://arxiv.org/pdf/2312.01479

ملخص

في هذا العصر الذي يتسارع فيه تطوّر الذكاء الاصطناعي على كافة الأصعدة، أصبحت الحاجة إلى حلول دبلجة صوتية عالية الجودة أمراً ضرورياً. غالبًا ما تنتج الأساليب التقليدية دبلجة صوتية روبوتية أو اصطناعية، مما يؤثر سلباً على تجربة المستخدم. يهدف هذا المشروع إلى تطوير تطبيق دبلجة صوتية يعتمد على الذكاء الاصطناعي لترجمة المحتوى الصوتي من اللغة الإنجليزية إلى اللغة العربية مع الحفاظ على خصائص الصوت الأصلي.

يعتمد النظام على خوارزميات ذكاء اصطناعي متقدمة لضمان إنتاج صوت مدبلج طبيعي وأصلي. يمكن للمستخدمين أيضاً اختيار أصوات بديلة، بما في ذلك أصوات المشاهير، أو تحميل تسجيلاتهم الصوتية المفضلة. بالإضافة إلى ذلك، يوفر التطبيق مخرجات نصية قابلة للتحرير لضمان الدقة، وإدارة المشاريع المدبلجة بفعالية من خلال إنشاء قوائم تشغيل وتنظيمها في مجلدات قابلة للتخصيص.

بهذا، يوفر النظام بيئة سهلة الاستخدام تمكن المستخدمين من إنتاج محتوى مدبلج عالي الجودة مع الحفاظ على نية وصوت المتحدث الأصلي.

الجامعة العربية الدولية

الجامعة العربية الدولية AIU جامعة سورية خاصة أُحدثت عام 2005، خططها الدراسية والوثائق الصادرة عنها معتمدة ومصدقة من قبل وزارة التعليم العالي في الجمهورية العربية السورية.

تعمل الجامعة على تحقيق الأهداف الآتية:

- إعداد جيل متميز من الخريجين الجامعيين القادرين على تلبية الحاجات النوعية للمجتمع والنهوض به.
- الإسهام في البحوث العلمية النظرية والتطبيقية التي تخدم أغراض التنمية الوطنية، ويتم العمل على حث الأساتذة والعاملين الأكاديميين على البحث العلمي والمشاركة في المؤتمرات والندوات التي تنظم الأبحاث.
- تحقيق الشراكة مع الجامعات العربية والأجنبية المرموقة بهدف التطوير والتحديث المستمرين للعمل الأكاديمي والقيام ببحوث علمية مشتركة.
- استقطاب الكفاءات الأكاديمية والبحثية المتميزة عن طريق توفير البيئة المناسبة لعملها.

الجامعة العربية الدولية من الجامعات السورية الأولى التي جرى تأسيسها وافتتاحها، وقد تمكنت من اجتذاب الكفاءات التعليمية والبحثية والإدارية المتميزة، إنشاء صرح متكامل من النواحي الأكاديمية والتنظيمية والإدارية. وتمكنت من تخريج كوادر من المبدعين والمتميزين من خالل توفير بيئة تعليمية ترتكز إلى مقومات نوعية ومادية فريدة منها:

- الخطط الدراسية الحديثة والمتطورة المستندة إلى نظام الساعات المعتمدة.

- الأطر التعليمية المنتقاة بعناية كبيرة.
- المختبرات العلمية الحديثة، ومختبر المكتبات اإللكترونية.
- المحفزات المادية والمعنوية للطلبة.
- تطبيق طرائق التدريس التفاعلي.
- التوجيه والإرشاد الأكاديمي والتربوي.
- مجموعة كبيرة من اتفاقيات التعاون العلمي مع جامعات محلية وإقليمية ودولية ذات سمعة مرموقة.
- اتفاقيات ومذكرات تفاهم متعددة مع العديد من مؤسسات المجتمع المدني.
- الحرم الجامعي الالئق والمزود بكافة المرافق العلمية والرياضية والترفيهية، والذي نشجعك على زيارته والتعرف على مزاياه.
- الأنشطة والأندية الطلابية بمختلف أنواعها: الرياضية والثقافية والعلمية واالجتماعية.

في الجامعة العربية الدولية سنوات الحياة الجامعية هي وقت الاستثمار في مستقبل الطالب. فالمعارف والخبرات التي يحصلها في قاعة المحاضرات والمختبرات ستساعده في تطوير ذاته، وستمنحه أسباب النجاح في التخصص الذي اختاره، والنشاط الطلابي الذي يمارسه سيساعده في توسيع أفقه، وفعاليات التدريب والأندية والرياضة ستمكنه من تطوير مواهبه، وربما تساعده في اكتشاف مواهب جديدة. ليستثمر وقته وذهنه وروحه في جامعتنا كي يجني فوائد عمله والوقت الذي كرسه في السنين القادمة. ونحن سوف نكون بجانب طلبتنا في كل خطوة على دربهم.

الجامعة العربية الدولية

كلية الهندسة المعلوماتية والاتصالات

مشروع التخرج بعنوان

الدبلجة الصوتية بالذكاء الاصطناعي

تم تقديمه إلى

قسم الهندسة المعلوماتية

تقديم

منى طحان **ماريا عماش**

ليث المزين

محمد صفوح غزال **جودي سويد**

بإشرا ف

المهندس محمد المصري **الدكتور طارق برهوم**

تموز 2024