FINAL YEAR PROJECT

ANIGEN

An automated Urdu Spokesperson

F22-091-R

FYP Team

Umair Afzal 19I-0517

Sameet Ikram 19I-0707

Umer Ahsan 19I-2184

Supervised By

Mr. Saad Salman

Co-supervised By

Dr. Mirza Omer Baig

FAST School of Computing

The National University of Computer and Emerging Sciences

Islamabad, Pakistan

2023

Students' Submission

Title: Final Year Project 2 Final-Evaluation Report

Anti-Plagiarism Declaration

This is to confirm that the following FYP (Final Year Project) report was produced under the:

Title: Anigen-Automated Urdu spokesperson

is the sole contribution of the authors and no part of this report has been copied as it is the basis (cut and paste), which is Plagiarism. All cited sources were used to support the FYP and were correctly credited. If a violation of this declaration is found, I/We shall be accountable and liable for any consequences.

Date: 4th June 2023

Student 1

Name: Umair Afzal

Student 2

Name: Sameet Ikram

Student 3

Name: Umer Ahsan

Supervisor (Faculty)

Name: Mr. Saad Salman

Author’s declaration

We thus state that this submission is completely our own unique work and that all quotations and research materials utilized in its preparation have been correctly cited.

Acknowledgment

We would like to thank our supervisor, Mr. Saad Salman, for his honest efforts and for giving us so much of his time. Our project has benefited greatly from his advice.

We would also like to thank the FYP committee for being a constant source of advice for us.

Last but not the least, we want to acknowledge our group members. Without the dedication and arduous work of the group members, this project would not be possible.

Executive Summary:

Getting input from the user in the form of Urdu text, image, and user's voice, Anigen will be able to generate a 3D animated avatar that will look like the user and its voice will be similar to that of the user's and it will speak in the Urdu language according to the input text.

Text to Speech (TTS) model will be used to convert the Urdu text into Urdu speech. The user can either type in the Urdu text or can upload it from a file.

The voice of the user will be used to clone the user's voice so that the result i.e. the 3D avatar, speaks in the user's voice.

The image of the user will be used to create a 3D avatar that will look like the user.

In the end, lip-sync will be applied on the avatar and all the above things will be integrated which will generate a video that will contain a 3D avatar that will speak in the Urdu language and will look similar to the user.

Table of Content

CHAPTER 1: INTRODUCTION

Nowadays, people use various software to make a single 3 animated video. For example, for creating an avatar, they use one software, lip-sync and facial expressions are done in another software, speech generation is done in another software, and then they need software to integrate all of these features. So, to save people from this problem, we will develop the Anigen. Anigen will help people create avatars that represent them, and people will only need to write the text. The app itself will do all the other work.

1.1: Problem Domain

Our final year project is based on the idea to generate a 3D video of a talking avatar in the Urdu language from Urdu text. We will first get the Urdu text and the user's image from the user as input. The text will pass through a TTS (Text to Speech) model to be converted into Urdu speech.

The user image that we took at the start will be used to create a user-like avatar. We will also clone the voice of the user by taking a minimum of a 5-minute recording from the user (We will not take the recording of the user if we already have the user's voice). We will pass the recording to our model which will clone the voice of the user.

Our project will apply the lip-sync technique on the avatar and will integrate the avatar with the speech generated and with the cloned voice. And in this way, a 3D video of the avatar will be generated.

1.2: Research Problem Statement

The former method for speech animation uses different software for voice-over 3D avatar creation, lip-sync, integration, etc. This wastes a lot of time. Further, you need knowledge and skills to use this software. So, we are automating a system to solve these problems.

CHAPTER 2: Literature Review

2.1: RESEARCH PAPERS

2.1.1

WAVENET: A GENERATIVE MODEL FOR RAW AUDIO (2016)

2.1.1.1 Summary

Text-to-speech techniques currently used concentrate on non-parametric creation (which combines brief audio signal segments from a large training set) and parametric production (in which a model generates acoustic features synthesised into a waveform with a vocoder). The audio generated by these approaches doesn't sound so natural, then a WaveNet was introduced in 2016 by DeepMind.

Unlike other text-to-speech systems at the time, WaveNet is a deep generative model that creates speech in the form of raw audio waveforms. Because it is an entirely probabilistic and autoregressive model in which every prior sample in the dataset influences the prediction of the new sample, it is based on the PixelCNN architecture. The following product of conditional probabilities is used to factorise a waveform with the joint probability x = x1,..., xt.:

Each audio sample at timestep xt is dependent on samples from all earlier timesteps. This model is created using dilated causal convolutions, which have a large receptive field and are useful for dealing with long-range temporal dependencies in probability calculations and the creation of raw audio.

Dilated Causal Convolution:

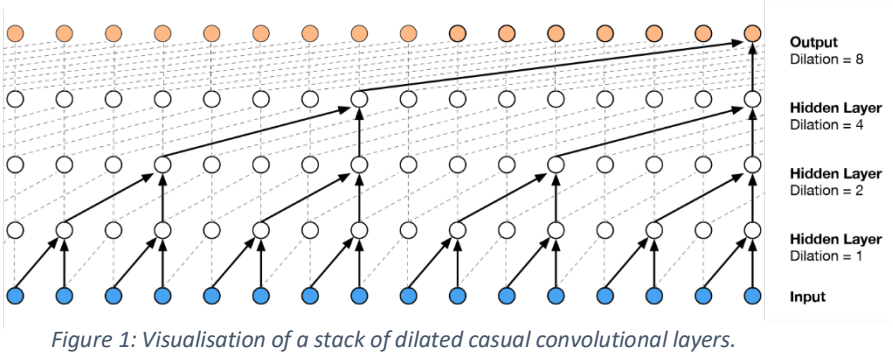


Figure 1: Visualisation of a stack of dilated casual convolutional layers.

In order to produce speech from text, we must condition WaveNet on both text and raw audio samples. This means that the network's predictions are based on both the text and the prior audio samples. The following equation can be used by WaveNet to predict the conditional distribution of the audio given an input h.

$$p(\vec{x}) = \prod_{t=1}^{T} p(x_t \mid x_1, \ldots, x_{t-1})$$

2.1.1.2 Critical Analysis

Strengths:

• Can produce raw speech signals with a subjective naturalness that has never been documented in the text-to-speech sector.

• Use dilated causal convolutions to handle long-range temporal dependencies.

• Multiple voices can be produced with a single model.

• Shows potential when used to produce other audio formats, like music.

Weaknesses:

- For WaveNet to synthesize one second of audio, it needs to operate for several minutes.
- However, producing the inputs to WaveNet (linguistic characteristics, anticipated log fundamental frequency (F0), and phoneme durations), which include sophisticated text-analysis tools and a substantial vocabulary, requires extensive domain expertise.
- No voice cloning on low resources.
- No one-shot or zero-shot learning.

2.1.2:

CHAR2WAV: END-TO-END SPEECH SYNTHESIS (17 Feb 2017)

2.1.2.1 Summary

The process of speech synthesis is split into two parts in conventional methods. The text is converted into linguistic features in the front end, which is the initial stage. The second stage, also referred to as the backend, creates the associated sound using input from the language features. In this model, the front end and back end have been merged, and the entire process has been learned.

There are two parts to Char2Wav: a reader and a neural vocoder. The reader, an encoder-decoder model with an attention mechanism, generates the acoustic properties of the vocoder from inputs such as text or phonemes. The conditional extension of SampleRNN that produces the raw audio waveforms from intermediate representations is the neural vocoder. It creates music based on the features that the reader produces as output. It assigns related audio samples to a series of vocoder features.

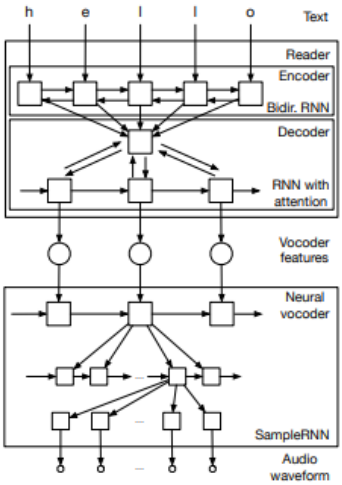


Figure 2: Char2Wav: An end-to-end speech synthesis model

For pre-training their alignment module, which includes F0, spectral envelope, and aperiodic parameters, Char2Wav uses vocoder features from the WORLD TTS system.

2.1.2.2 Critical Analysis

Strengths:

- Learns to produce audio directly from the text.
- Integrate the front end and the back end and learn the whole process end-to-end.
- Eliminates the need for expert linguistic knowledge.

Weaknesses:

- Vocoder features as intermediate representations.
- Separate training of reader and vocoder and cannot be trained parallelly.
- No direct alignment between characters and audio samples.
- No voice cloning on low resources.
- No one-shot or zero-shot learning.

2.1.3:

DEEP VOICE: REAL-TIME NEURAL TEXT TO SPEECH (25 Feb 2017)

2.1.3.1 Summary

A production-quality TTS system must have real-time inference; otherwise, the system cannot be used for the majority of applications. Although WaveNet's inference is a challenging computational challenge, it can nevertheless produce speech that is nearly human-like and can therefore be employed in production systems.

Deep Voice is modeled after conventional text-to-speech pipelines and uses a similar architecture, but all of the components are replaced by neural networks, and the features are simplified. First, the text is converted to phonemes, and then the linguistic features are translated into speech using an audio synthesis model. They only have phonemes with stress annotations, phoneme lengths, and fundamental frequency, in contrast to earlier research.

There are five main parts to the deep voice system:

- Using the ARPANET database as a base, the grapheme-to-phoneme model translates text to phonemes.
- In the audio file, the segmentation model finds the phoneme boundary.
- The length of each phoneme in a sequence is predicted by the phoneme duration model.
- Using the duration of each phoneme, the fundamental frequency model calculates the fundamental frequency F0 of each phoneme.
- The phoneme, phoneme duration, and F0 are used as local conditioning input features by the WaveNet audio synthesis model to produce the final utterances.

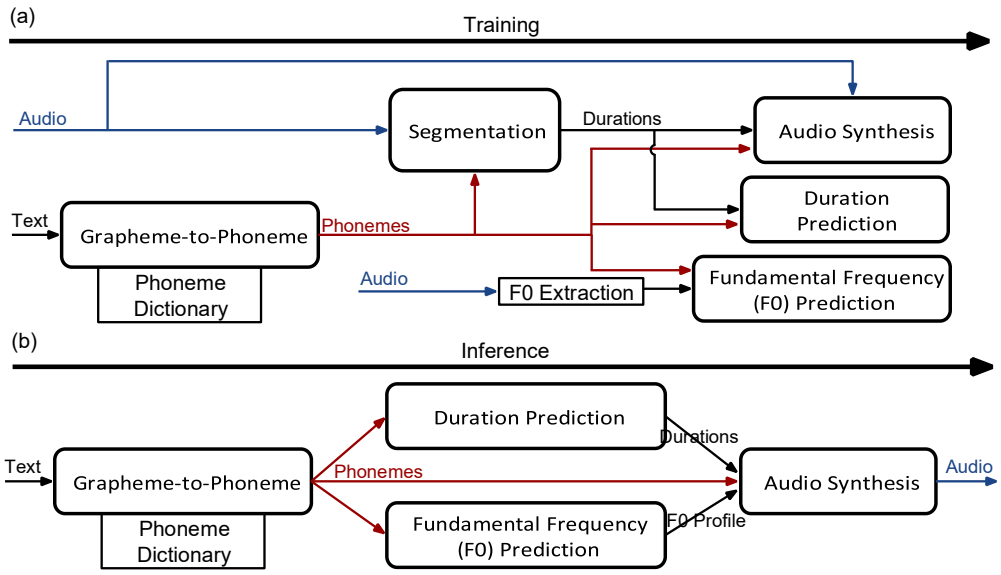


Figure 3: System diagram depicting a) training procedure and b) inference procedure

Since the segmentation model just needs to mark the phoneme boundaries in the training audio data, it is not employed in the inference process.

Deep Voice provides a tuneable trade-off between speed of synthesis and quality of audio and can synthesize sounds in fractions of a second. On the other hand, prior WaveNet results need many minutes of runtime to synthesize a single second of audio.

2.1.3.2 Critical Analysis

Strengths:

- is entirely independent; no pre-existing TTS system is necessary for training a new Deep Voice system. A dataset of brief audio samples and associated written transcripts can be used for training from scratch.
- Deep Voice offers a tuneable trade-off between synthesis speed and audio quality and can synthesize sounds in fractions of a second.
- The number of hand-engineered characteristics in Deep Voice is kept to a minimum; in contrast to earlier work, our sole features are phonemes with stress annotations, phoneme durations, and fundamental frequency (F0).

Weaknesses:

- There is no evidence that its naturalness rivals that of human speech.
- Still use of hand-engineered features thus requires domain knowledge.
- Use of WaveNet as a vocoder slows the process.
- No voice cloning on low resources.
- No one-shot or zero-shot learning.

2.1.4

Natural TTS Synthesis by Conditioning WaveNet on Mel Spectrogram Predictions (16 Dec 2017)

2.1.4.1 Summary

For a long time, the cutting-edge method was concatenative synthesis or the joining of small units of previously recorded waveforms. Concatenative synthesis, which directly constructs smooth trajectories of speech features to be synthesized by a vocoder, encountered many of the same boundary artifact issues as statistical parametric voice synthesis. The audio that these algorithms generate, however, frequently sounds muffled and artificial when compared to human speech. Figure 4 depicts the two parts of the proposed system: (1) From an input character sequence, a recurrent sequence-to-

sequence feature prediction network with attention predicts a series of Mel spectrogram frames. (2) Time-domain waveform samples are generated using a modified version of WaveNet in dependence on the expected Mel spectrogram frames.

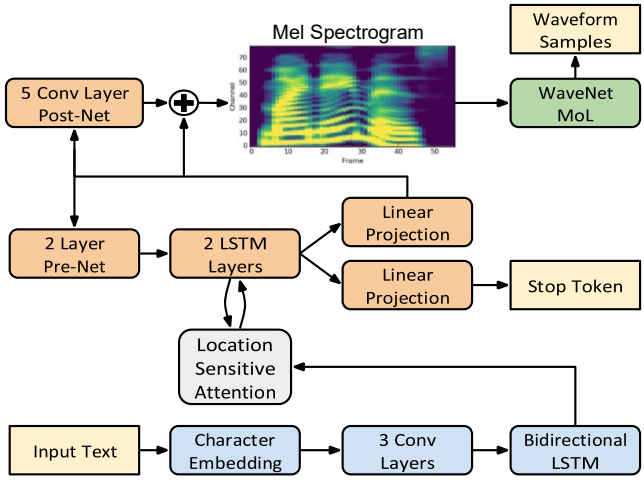


Figure 4: Block diagram of the Tactron 2 system architecture.

A fully neural TTS system predicts Mel spectrograms using a modified WaveNet vocoder and a sequence-to-sequence recurrent network. The resulting system is capable of synthesizing speech with prosody and audio quality comparable to WaveNet. This approach may be trained directly from data without the use of intricate feature engineering, and it provides a state-of-the-art sound quality that is comparable to that of natural human speech.

2.1.4.2 Critical Analysis

Strengths:

- Direct training using speech waveforms and normalized character sequences.
- selected a weak acoustic representation: To connect the two parts, use Mel frequency spectrograms. We train the two components independently utilizing a representation that is easily computed from time-domain waveforms.
- The suggested approach greatly surpasses all other TTS systems and generates a MOS that is similar to the audio from the source material.

Weaknesses:

- Because it is autoregressive, it cannot be parallelized or trained end-to-end.
- Required the intermediate text representation i.e. phonemes.
- Use of WaveNet for audio wave generation slows the process.
- No voice cloning on low resources.
- No one-shot or zero-shot learning.

2.1.5:

Conditional Variational Autoencoder with Adversarial Learning for End-to-End Text-to-Speech (11 Jun 2021)

2.1.5.1 Summary

Apart from text pre-processing like text normalization and phonemization, the TTS system pipeline has been streamlined to two stages of generative modeling due to the rapid growth of neural networks. Realistic speech synthesis has been demonstrated by neural network-based autoregressive TTS systems, but their sequential generation approach makes it challenging to take full advantage of contemporary parallel processors. Although parallel TTS systems have advanced, two-stage pipelines are still a challenge since they need sequential training.

This research presents a parallel end-to-end TTS approach that produces audio that sounds more realistic than existing two-stage models. The two modules of TTS systems are linked via a variational autoencoder latent variable to facilitate effective end-to-end learning. Adversarial training on the waveform domain and normalizing flows to conditional prior distribution was used to increase the expressive capacity of this method. They also suggest a stochastic duration predictor creates speech with a variety of rhythms from input text to address the one-to-many challenge. Compared to the finest publicly accessible TTS system, our technique produces speech that sounds more natural and has a greater sampling efficiency.

A decoder, posterior encoder, discriminator, prior encoder, and stochastic duration predictor make up the model's overall architecture. Only training uses the posterior encoder and discriminator.

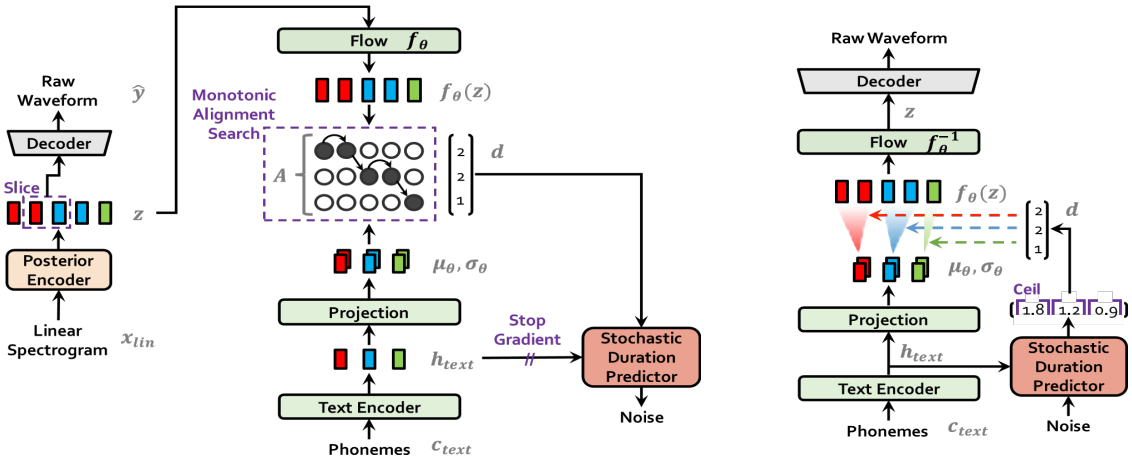


Figure 5 (a) Training Procedure *(b) Inference Procedure*

2.1.5.2 Critical Analysis

Strengths:

- a simultaneous end-to-end TTS approach that produces audio that sounds more realistic than the two-stage models used today.
- One-to-many relationships allow for the many pronunciations of a single text input in a variety of tones and rhythms.
- delivers a MOS similar to ground truth and surpasses the greatest publicly accessible TTS systems.
- captures differences in voice that are not reflected by text.
- learns to produce raw waveforms from text alone, without the need for additional input conditions
- instead of calculating the loss, searches for the ideal alignment using the dynamic programming technique known as MAS.

Weaknesses:

- No one-shot or zero-shot learning.
- Slow training process.
- Required the intermediate text representation i.e. phonemes for quality results.
- No expressive speech synthesis.

2.1.6:

COMPARISON OF TEXT-TO-SPEECH MODELS:

Model Name	Release Date	Proposed Approach	MOS Value				Datasets
			Single Speaker		Multi Speaker		
			GT	MOS	GT	MOS	
WaveNet	19 Sep 2016	N-gram Probabilistic model with a dilated causal convolutional network	4.55 ±0.075, 4.21 ±0.071	4.21 ±0.081, 4.08 ±0.085		-	VCTK, LJ Speech, Mandarin Chinese
Deep Voice	7 Mar 2017	Replacing all components of the TTS pipeline with neural networks	4.34 ±0.18	3.94 ±0.26	-	-	Blizzard 2013
Char2Wav	16 Apr 2017	Encoder-Decoder architecture with attention and SampleRNN as a vocoder	-	-	-	-	VCTK
Tacotron 2	16 Feb, 2018	Modified WaveNet as a vocoder and Encoder-Decoder architecture with attention.	4.582 ±0.053	4.526 ±0.066	-	-	LJ Speech
VITS	11 Jun 2021	Conditional variational autoencoders with adversarial learning	4.46 ±0.06	4.43 ±0.06	4.38 ±0.07	4.38 ±0.06	LJ Speech VCTK

Table 1: Comparison of text-to-speech models

2.2: RESEARCH COMPONENTS

2.2.1 Dataset Generation

Many of the text-to-speech models need their data in form of short clips of 6 to 10 seconds. Further, the transcript of the clip along with the speaker's information is also required. The audio should be recorded in a quiet place so that there is no other background sound. Further, the recorded audio should be in one tone and there should be no pause in speech while recording.

2.2.1 Speech Generation

The majority of text-to-speech models' preprocessing requires a phonemizer as they don't take text directly but instead use the phonemes of that language that are written in IPA standards. There is no open-source phonemizer for the Urdu language that makes phonemes of Urdu according to the IPA standard.

2.2.3 Low Resource

All the models that have been made so far are for languages that are rich in resources. There are no models for a low-resource language like Urdu. Therefore, we will make such a model that will well on the low resources as well.

CHAPTER 3: SRS

3.1 List of features

- Creation of a 3D avatar similar to the input image
- Generation of 3D video
- Selection of template
- Conversion of Urdu text to speech
- Cloning of user's voice

3.2 Functional Requirements:

- User will have the option to build a 3D avatar.
- User will have the option to create a 3D video.
- User will be able to select a template either from his/her device or from the already available templates
- User will be able to convert Urdu text into speech
- User will be able to clone his/her voice

3.3 Quality Attributes

- Efficiency
- Simplicity
- Availability
- Performance

3.4 Non-functional Requirements

- The system shall be available 24/7
- Each page shall load within 2 seconds
- The system shall be able to handle about 10000 user requests per second

CHAPTER 4: Proposed Approach

In this study, we present a resemblant end-to-end TTS system that produces audio that sounds more realistic than existing two-stage models. We link two TTS system modules together using latent variables and a variational autoencoder (VAE) (Kingma & Welling, 2014) to facilitate effective end-to-end learning. We add normalising flows to our conditional prior distribution and adversarial training on the waveform domain to enhance the expressive capacity of our approach and synthesis high-quality speech waveforms. TTS systems need to explain the one-to-many relationship in which text input can be spoken in various ways with different variants in addition to producing fine-grained audio (e.g., pitch and duration). We also suggest a stochastic duration predictor to create speech with various rhythms from input text in order to address the one-to-many dilemma. Our method captures voice changes that cannot be represented by text by using the stochastic duration predictor and uncertainty modelling over latent variables.

Compared to the best publicly available TTS system, Glow-TTS (Kim et al., 2020) with HiFiGAN, our technique produces speech that sounds more realistic and has a higher sampling efficiency (Kong et al., 2020). We release the source code and our demo page to the public.

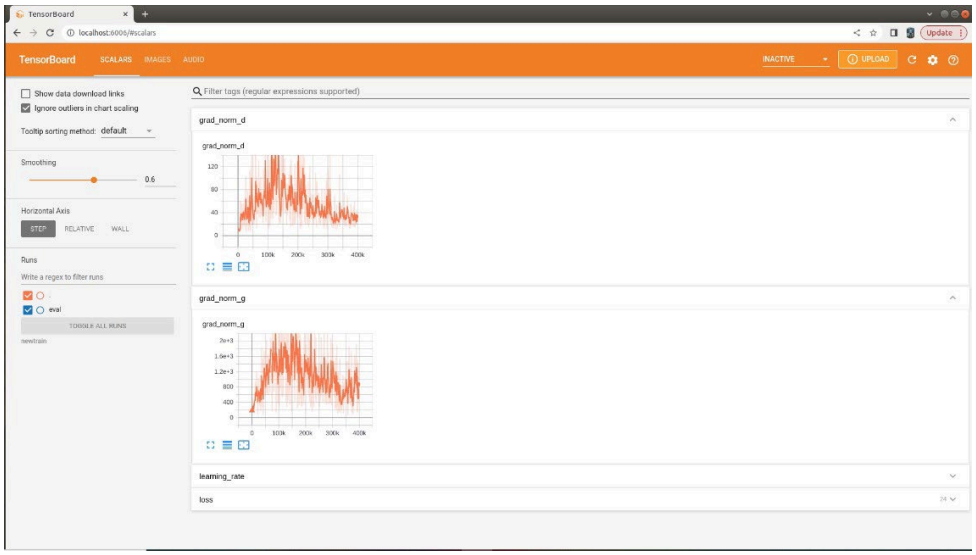
CHAPTER 5: IMPLEMENTATION

To create the text to speech system for Urdu, we first need to convert the text into the phonemes which provides more broader representation of words and align better with the sounds instead of raw text. Then the text to speech model, which is YourTTS architecture based on variational learning, is fine tuned on text and corresponding Mel spectrograms of our Urdu dataset. The original pretrained speaker encoder which is trained on voices of thousands of speakers is used to generate the speaker embeddings for voice cloning purposes because of low amount of dataset. To generate the raw speech signals, the hifi-gan vocoder is used along with the text to Mel spectrogram and speaker encoder system. The whole pipeline is trained in parallel end-to-end manner.

CHAPTER 6: Results

5.1 Gradient:

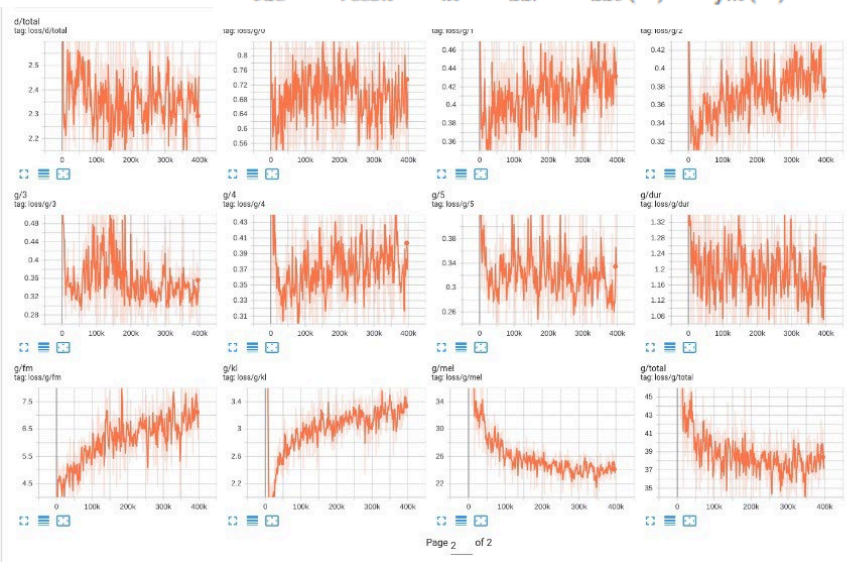
We have two gradients. One is gradient G and other is Gradient D. Gradient G is the generator network that generates the audio in the text. Gradient D is the discriminator network that differentiate how much is the generative audio different from actual audio.

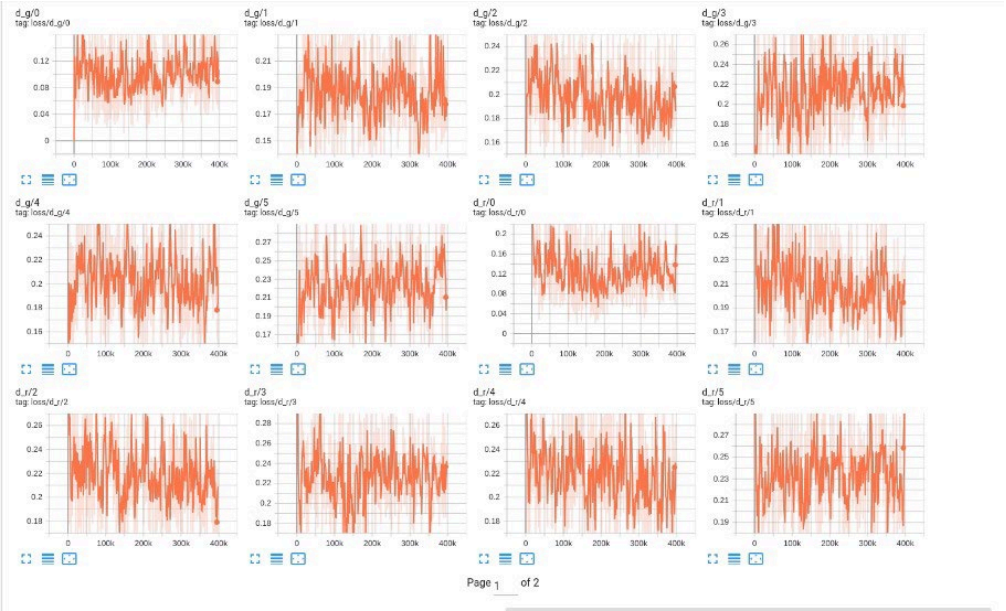


5.2 LOSS:

There are different types of losses in Generator network. These are: Mel loss, kl loss, duration loss, adversarial loss. All of these losses are shown in the graphs given below. The graphs for all other losses are decreasing except for kl loss that is increasing. This is the result of adversarial training in which one loss increases and other decreases. The results are shown in the images given below. The loss equation is:

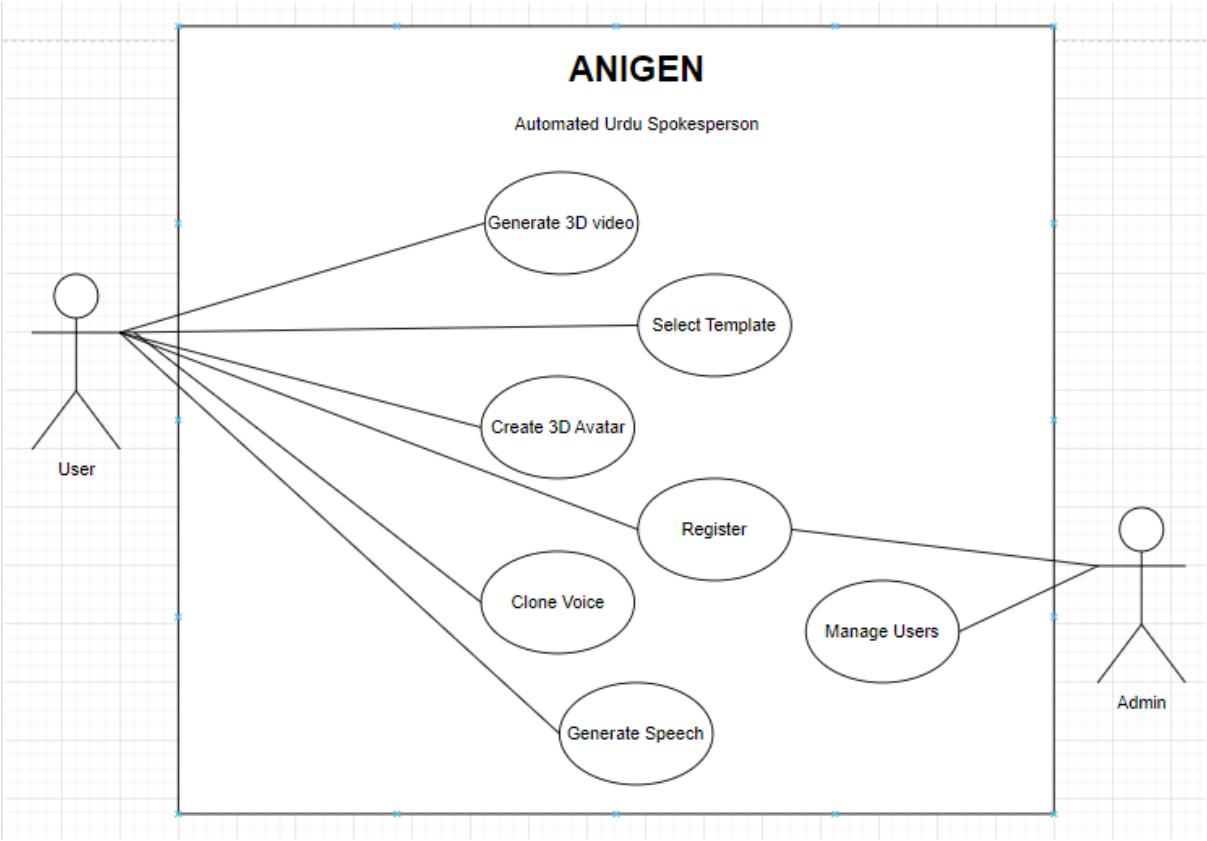
$$L_{vae} = L_{recon} + L_{kl} + L_{dur} + L_{adv}(G) + L_{fm}(G)$$





CHAPTER 7: UML DIAGRAMS

6.1 USE CASE DIAGRAM:



6.2 HIGH-LEVEL USE CASE:

Register an account

Use Case	Register an Account
Actors	User, System
Type	Primary, offstage
Description	The user opens the website. Then selects the option "Register an account". The user inputs his/her details along with the password. The system successfully registers the user.

Select Template

Use Case	Select Template
Actors	User, System
Type	Primary, offstage
Description	The user opens the website. Then selects the option of "Template" from the navbar. The user tells whether he wants to select a predefined template or he wants to add a picture from his device. The user selects the required template and the template is set at the back of his video.

Generate Speech

Use Case	Generate Speech
Actors	User, System
Type	Primary, offstage
Description	The user enters text in Urdu language and the system converts that text into Urdu speech.

Clone Voice

Use Case	Clone Voice
Actors	User, System
Type	Primary, offstage
Description	The user inputs a 5-minute voice as input and the system clones his voice and now, any speech can be produced by the system in the user's voice.

Generate 3D video

Use Case	Generate 3D video
Actors	User, System
Type	Primary, offstage
Description	The user opens the website. Then selects the option of generating a video. He provides the voice, text, template, and avatar to the system and the system generates a 3D video for him.

Create 3D avatar

Use Case	Create 3D avatar
Actors	User, System
Type	Primary, offstage
Description	The user opens the website. Then selects the option of creating an avatar. He provides his image to the system and the system generates a 3D Avatar for him.

Manage Users

Use Case	Manage Users
Actors	User, Administration, System
Type	Primary, Primary, and offstage
Description	The admin can add new users, can delete an existing user, or can update the details of an existing user.

6.3 EXTENDED USE CASE:

Register an account

Use Case Name	Register an Account	
Scope	Anigen Automated Urdu spokesperson	
Level	User's intended outcome	
Primary Actor	User	
Stakeholders and Interests	User	
Preconditions	The user is currently signed in.	
Success Guarantee	The user will be able to successfully register his/her account.	
Main Success Scenario	User's Actions 1: The user opens the website and clicks the register button. 3: The user submits the form after entering his name, phone number, username,	System's Response 2: The system asks for some details from the user. 4: The system successfully registers the user.

	password, and other information.	
Extensions	If the password is less than 5 characters then an error message will occur.	

Select Template

Use Case Name	Select Template	
Scope	Anigen Automated Urdu spokesperson	
Level	User's intended outcome	
Primary Actor	User	
Stakeholders and Interests	User	
Preconditions	The user is currently signed in.	
Success Guarantee	The user will be able to successfully select templates either from his/her device or from already available templates.	
Main Success Scenario	User's Action 1: The user selects the option of "select template". 3: The user selects the option accordingly. 5: The user selects the desired template.	System's Response 2: The system gives the user the option to select a template from the user's device or built-in templates. 4: The system shows various templates to the user. 6: The system applies the template at the back of the video.
Extensions	None	

Generate Speech

Use Case Name	Generate speech	
Scope	Anigen Automated Urdu spokesperson	
Level	User's intended outcome	
Primary Actor	Users	
Stakeholders and Interests	Users	
Preconditions	The user is currently signed in.	
Success Guarantee	The user will be able to generate Urdu speech from Urdu text.	
Main Success Scenario	User's Actions 1: The option to produce speech is chosen by the user. 3: The user enters Urdu text in the text box and enters submit button.	System's Response 2: The computer system offers a text box that the user can fill up. 4: The system generates Urdu speech from the Urdu text provided.
Extensions	The entered text will have a limit and if the user exceeds that limit then an error message will be shown to him.	

Clone voice

Use Case Name	Clone Voice
Scope	Anigen Automated Urdu spokesperson
Level	User's intended outcome
Primary Actor	User
Stakeholders and Interests	User

Preconditions	The user is currently signed in.	
Success Guarantee	The user will be able to successfully clone his voice.	
Main Success Scenario	User's Actions 1: The user selects the option to clone voice. 3: The user inputs his/her recorded voices.	System's Response: 2: The system asks the user to input a voice recording of at least 5 minutes. 4: The system clones the voice of the user.
Extensions	If the user inputs a voice in less than 5 minutes then an error message will be shown.	

Generate 3D video

Use Case Name	Generate 3D video	
Scope	Anigen Automated Urdu spokesperson	
Level	User's intended outcome	
Primary Actor	User	
Stakeholders and Interests	User	
Preconditions	1. The user is currently signed in. 2. The user has cloned his/her voice. 3. The user has made his 3D avatar.	
Success Guarantee	The user will be able to successfully generate a 3D video.	
Main Success Scenario	User's Action 1: The user selects "Generate Video" from the menu.	System's Response 2: The system asks the user to enter the text in Urdu language. 4: The system combines the speech, cloned voice,

	3 The user enters the text.:	avatar, and template along with lip-syncing and generates a 3D video.
Extensions	1. If the avatar is not already created then an error message will pop up 2. If the voice is not cloned then an alert message will be shown to the user.	

Create 3D avatar

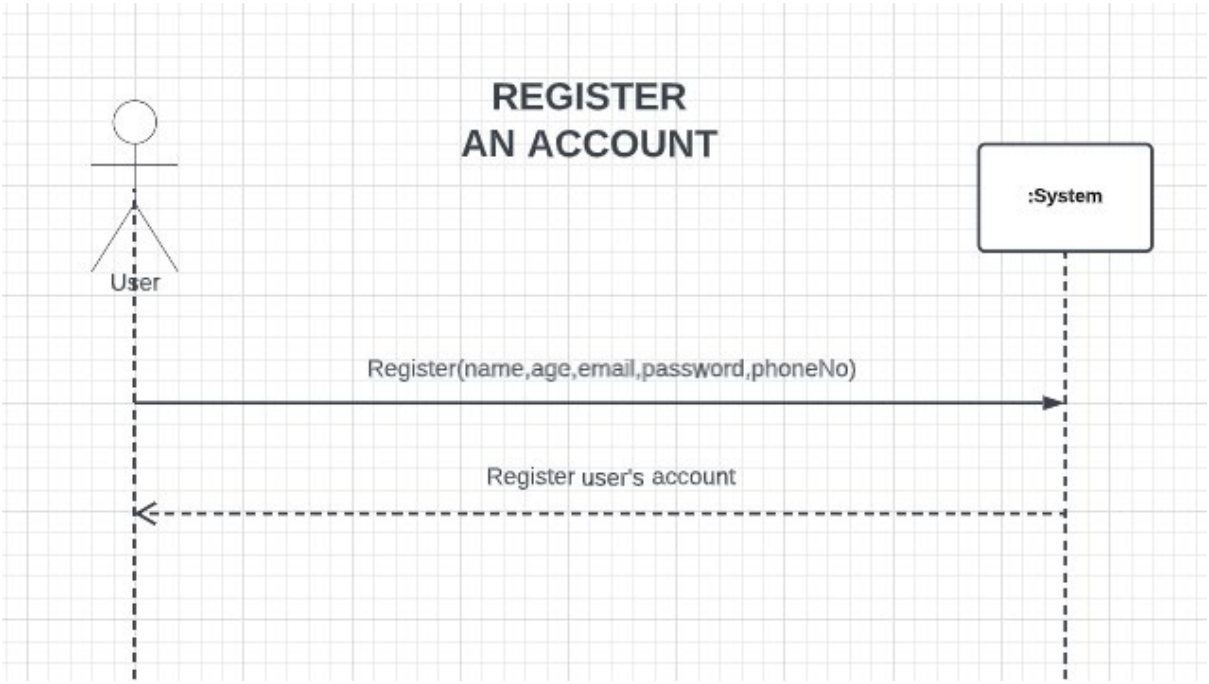
Use Case Name	Create 3D avatar	
Scope	Anigen Automated Urdu spokesperson	
Level	User's intended outcome	
Primary Actor	User	
Stakeholders and Interests	User	
Preconditions	The user is currently signed in.	
Success Guarantee	The user will be able to create a 3D avatar according to the picture provided.	
Main Success Scenario	User's Action 1: The "create Avatar" button is clicked by the user. 3: The user submits a picture.	System's Response 2: The user's image is requested by the computer system. 4: The system creates a 3D avatar from the picture provided by the user.
Extensions	None	

Manage users

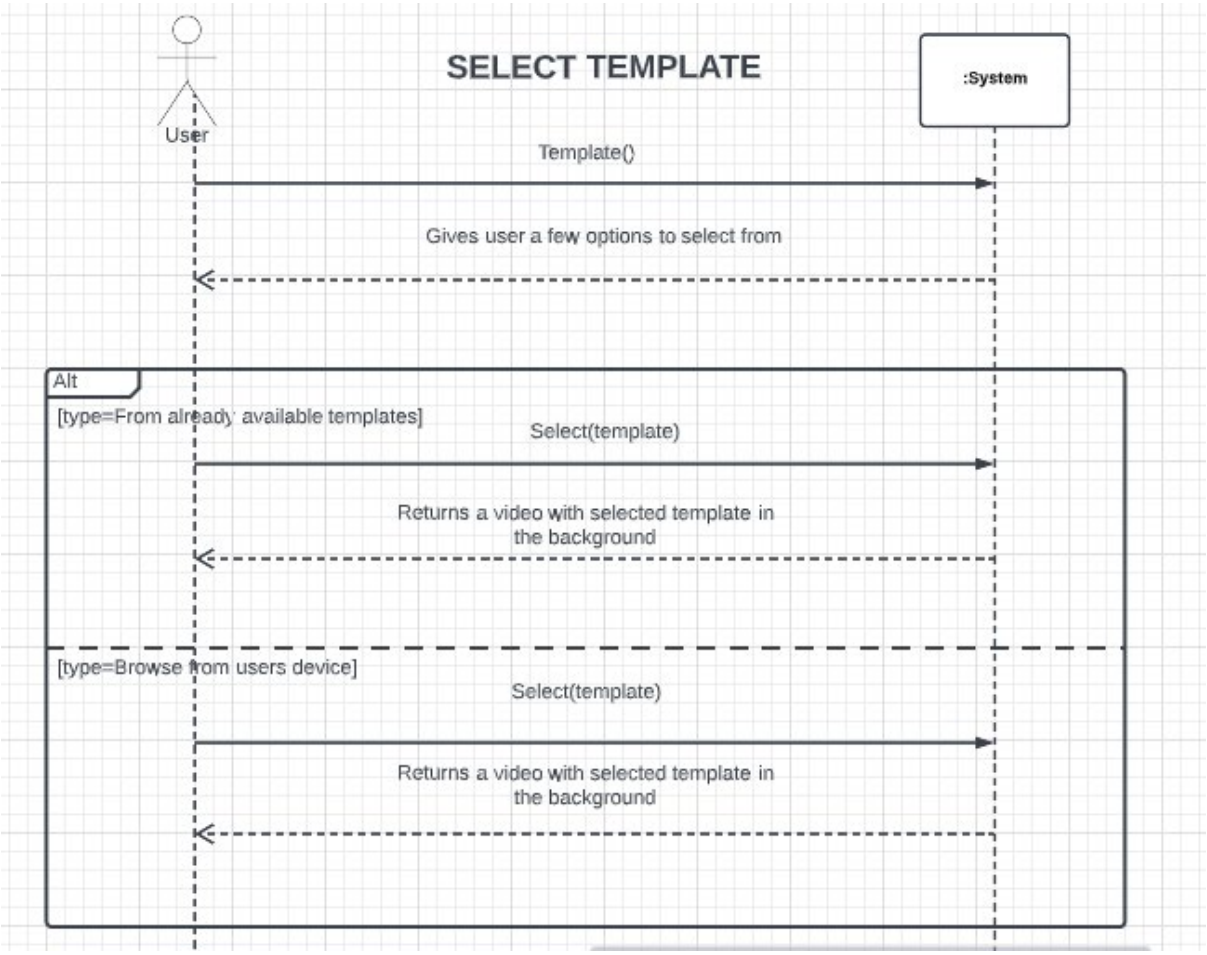
Use Case Name	Manage Users	
Scope	Anigen-Automated Urdu spokesperson	
Level	Administrator's goal	
Primary Actor	Administrator	
Stakeholders and Interests	Users, Administrator	
Preconditions	The user should already have an account if it needs to be updated or deleted.	
Success Guarantee	The user's account will be added/ removed/ updated.	
Main Success Scenario	User's Actions 1: The option of managing users is chosen by the administrator. 3: The administrator selects the option of "view all users", "delete user" or "add a new user".	System's Response 2: The computer system presents the administrator with a number of options. 4: The system shows the details to the admin or deletes the account of the user or adds a new user according to the choice selected by the user.
Extensions	If there is no user yet registered in the system, then the system shows a pop-up message to the admin that no user is yet present.	

6.5 SYSTEM SEQUENCE DIAGRAMS:

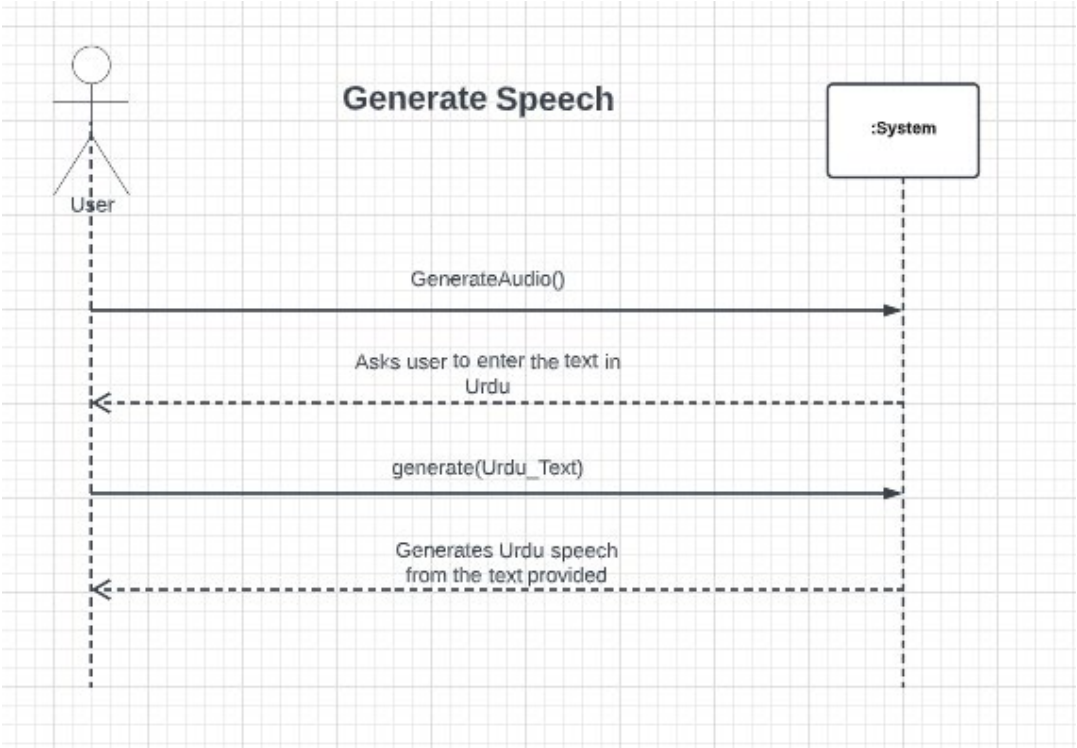
Register an Account



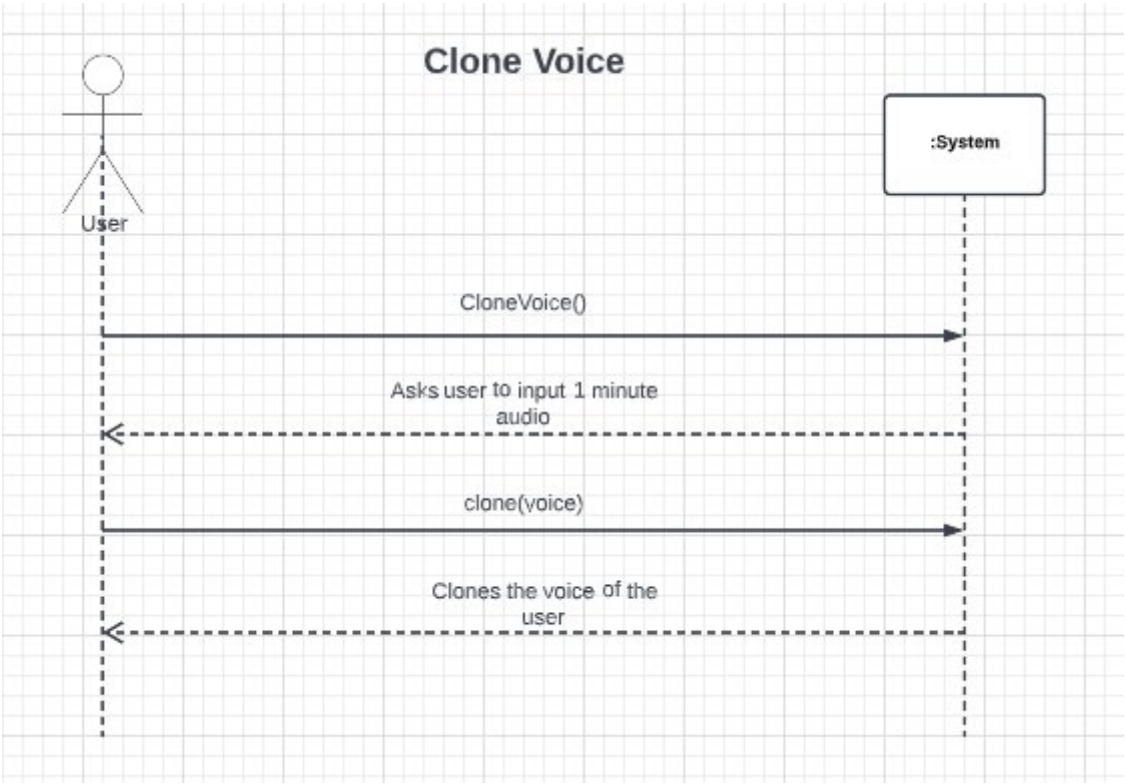
Select Template



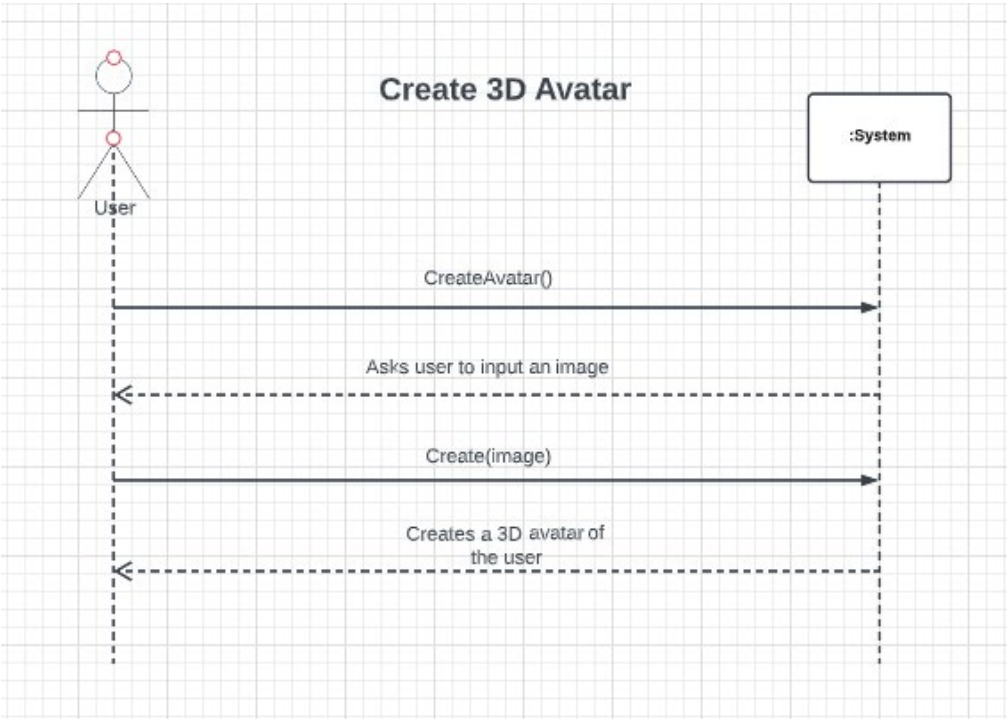
Generate Speech



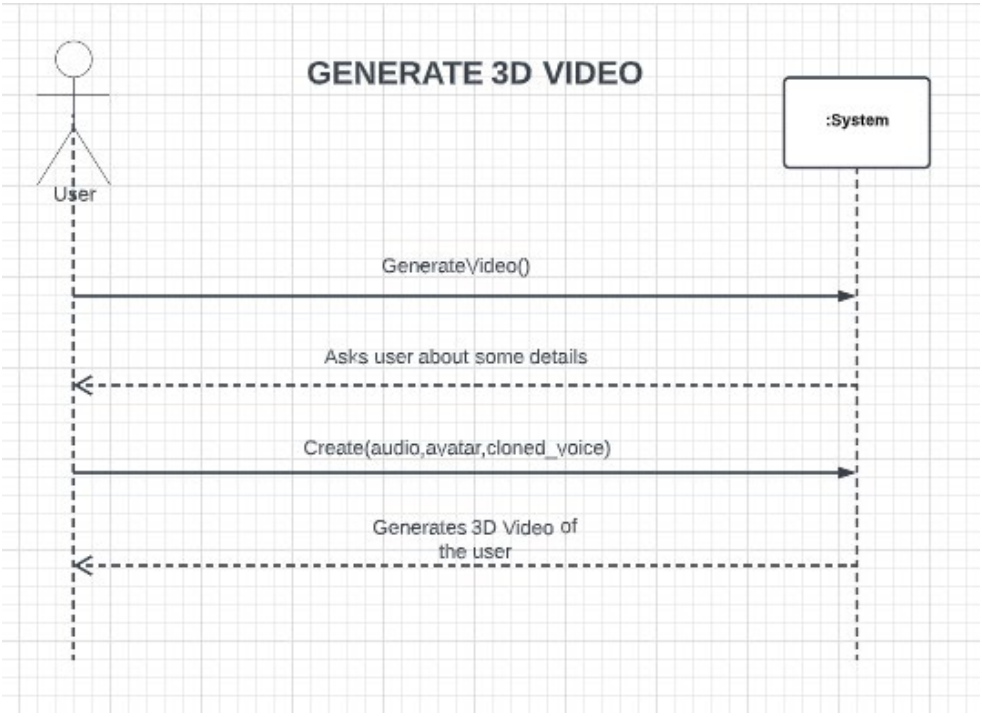
Clone Voice



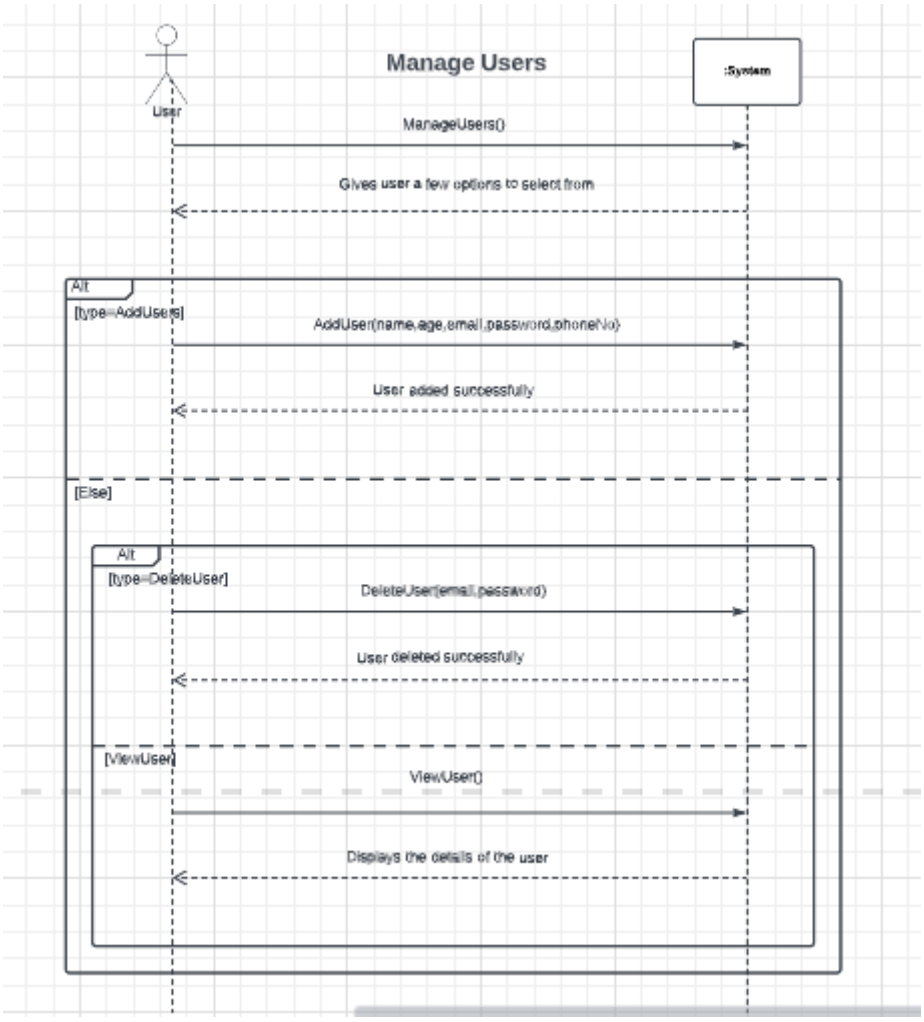
Create 3D Avatar



Generate 3D video



Manage Users



CONCLUSION

In conclusion, our work introduces an innovative end-to-end Text-to-Speech (TTS) system that surpasses current two-stage models in generating natural-sounding audio. By incorporating a variational autoencoder (VAE) and idle variables, we establish an effective end-to-end approach. We improve the quality of synthesized speech waveforms through homogenizing overflows and adversarial training. Our TTS system addresses the challenge of one-to-numerous relationships by employing a stochastic duration predictor to synthesize speech with varying measures.
Our end-to-end TTS system represents a significant advancement, offering high-quality and diverse speech synthesis. The incorporation of VAE, stochastic duration predictor, and query modelling over idle variables enables the capture and synthesis of speech variations beyond the limitations of the input text. This work opens up new avenues for enhancing the quality and expressiveness of TTS systems, facilitating the development of more realistic and engaging synthesized speech applications.

REFERENCES

1. Oord, Aaron van den, et al. "Wavenet: A generative model for raw audio." *arXiv preprint arXiv:1609.03499* (2016).
2. Arık, Sercan Ö., et al. "Deep voice: Real-time neural text-to-speech." *International Conference on Machine Learning*. PMLR, 2017.
3. Sotelo, Jose, et al. "Char2wav: End-to-end speech synthesis." (2017).
4. Shen, Jonathan, et al. "Natural tts synthesis by conditioning wavenet on Mel spectrogram predictions." *2018 IEEE international conference on acoustics, speech, and signal processing (ICASSP)*. IEEE, 2018.
5. Kim, Jaehyeon, Jungil Kong, and Juhee Son. "Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech." *International Conference on Machine Learning*. PMLR, 2021.