VOICE CLONING

Capstone Project Report

Submitted by:

(101610091)Taranjeet Singh

(101603354)Tarun Dani

(101603352)Tanuj Kukreja

BE Fourth Year, COE

CPG No. 20

Under the Mentorship of

Dr. Raman Kumar Goyal

Assistant Professor

THAPAR INSTITUTE
OF ENGINEERING & TECHNOLOGY
(Deemed to be University)

Computer Science and Engineering Department

Thapar Institute of Engineering and Technology, Patiala

May, 2020

TABLE OF CONTENTS

Use Case Diagram

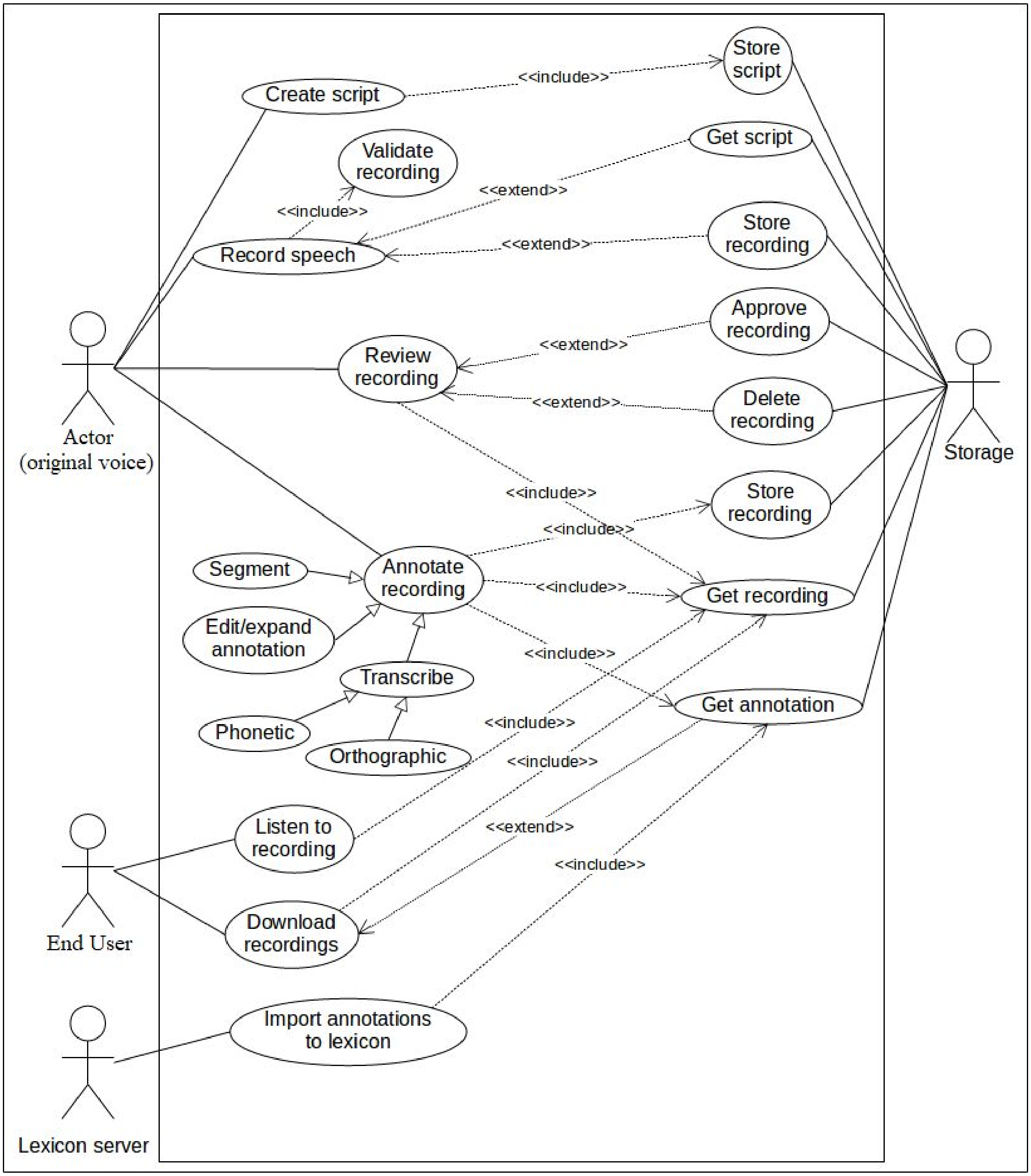
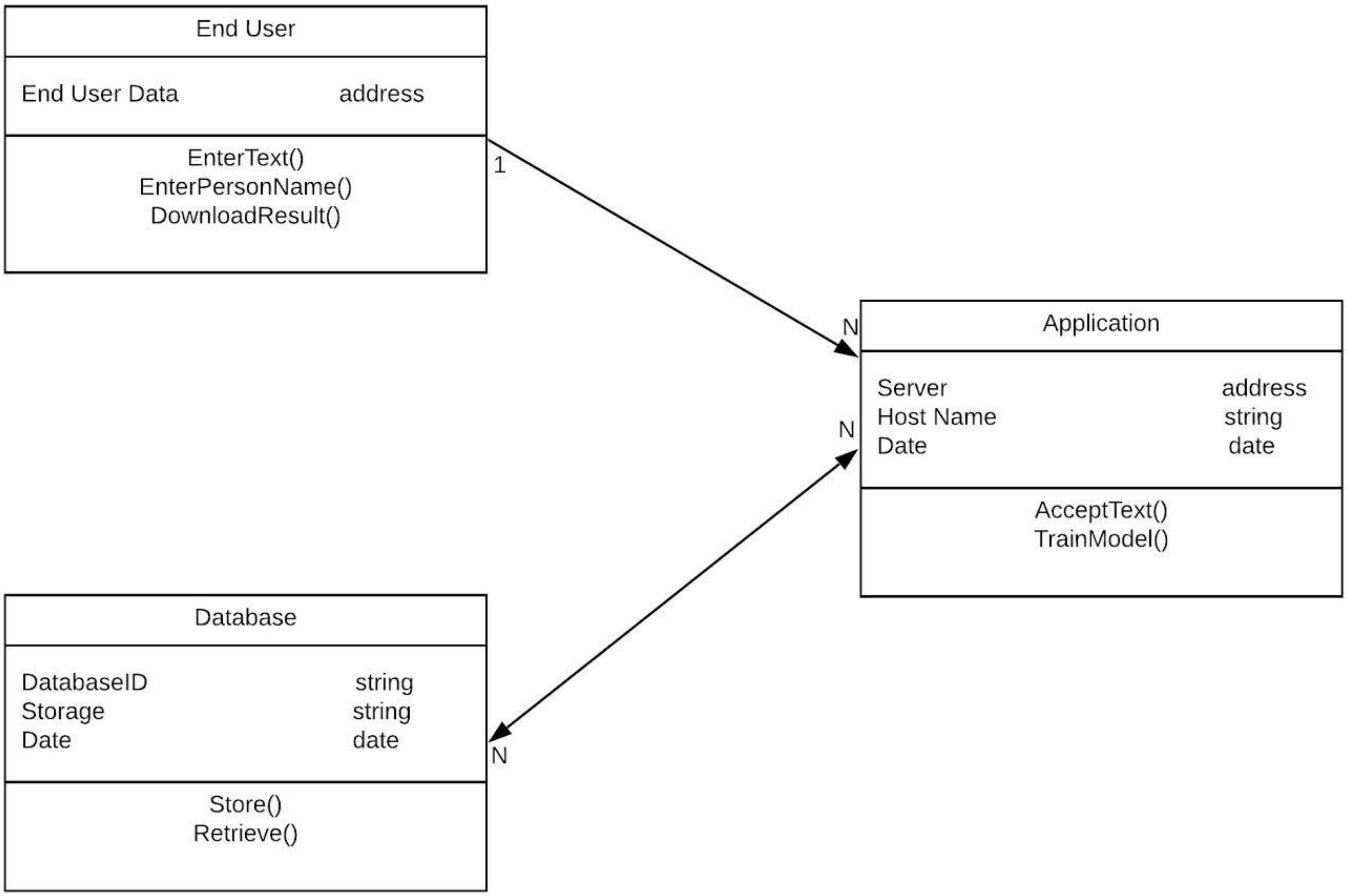
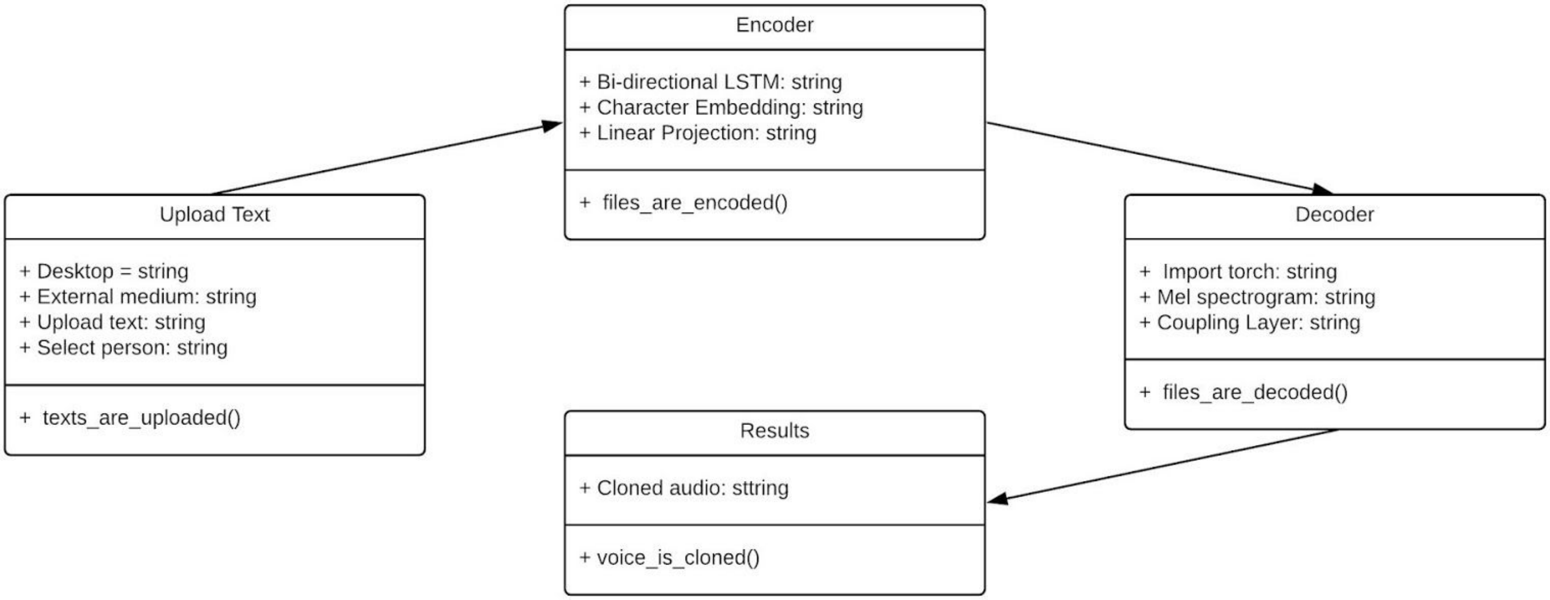


Fig.

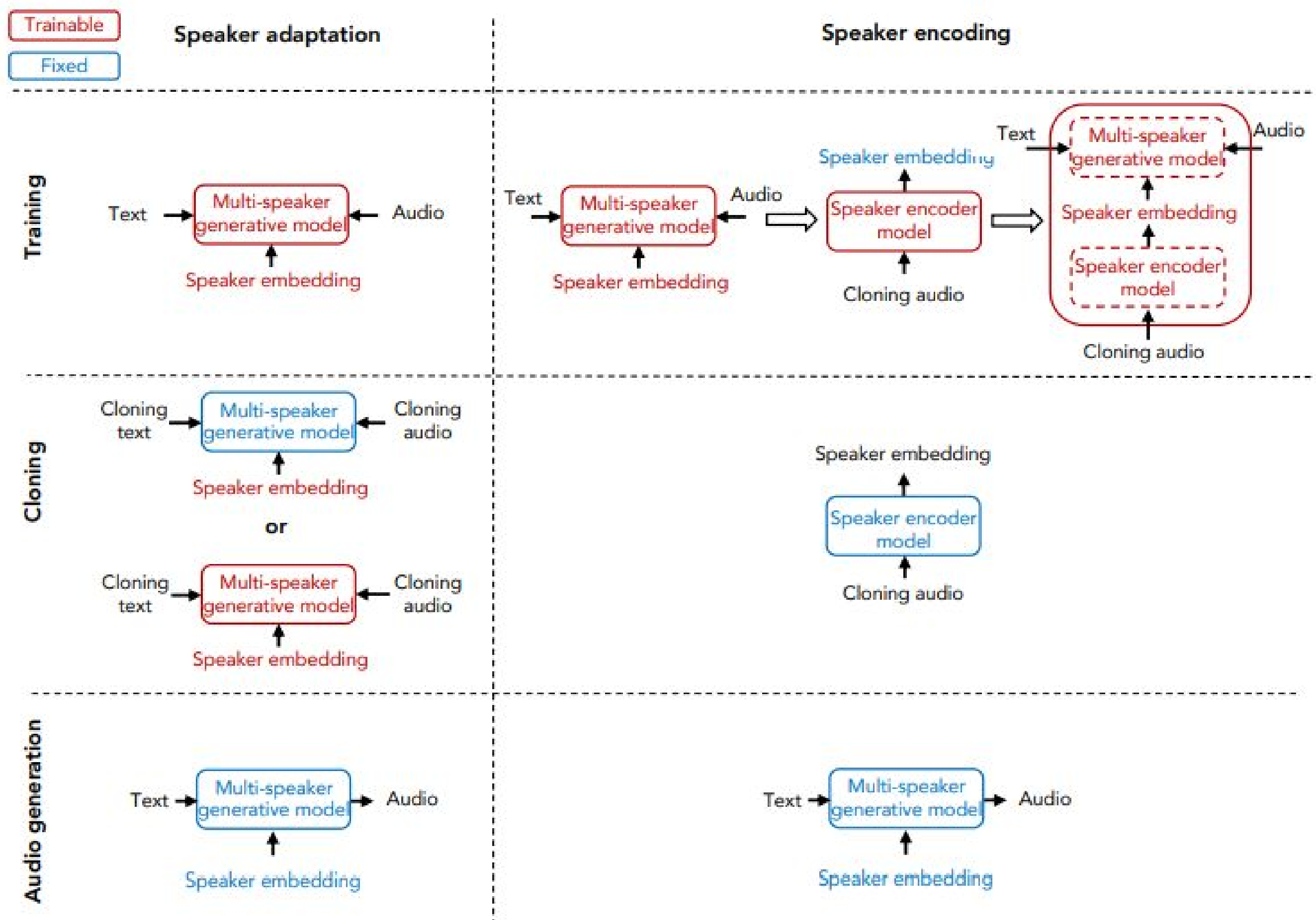
E-R Diagram


End User
End User Data address
EnterText()
EnterPersonName()
DownloadResult()
1
N
Application
Server address
Host Name string
Date date
AcceptText()
TrainModel()
N
Database
DatabaseID string
Storage string
Date date
Store()
Retrieve()
N

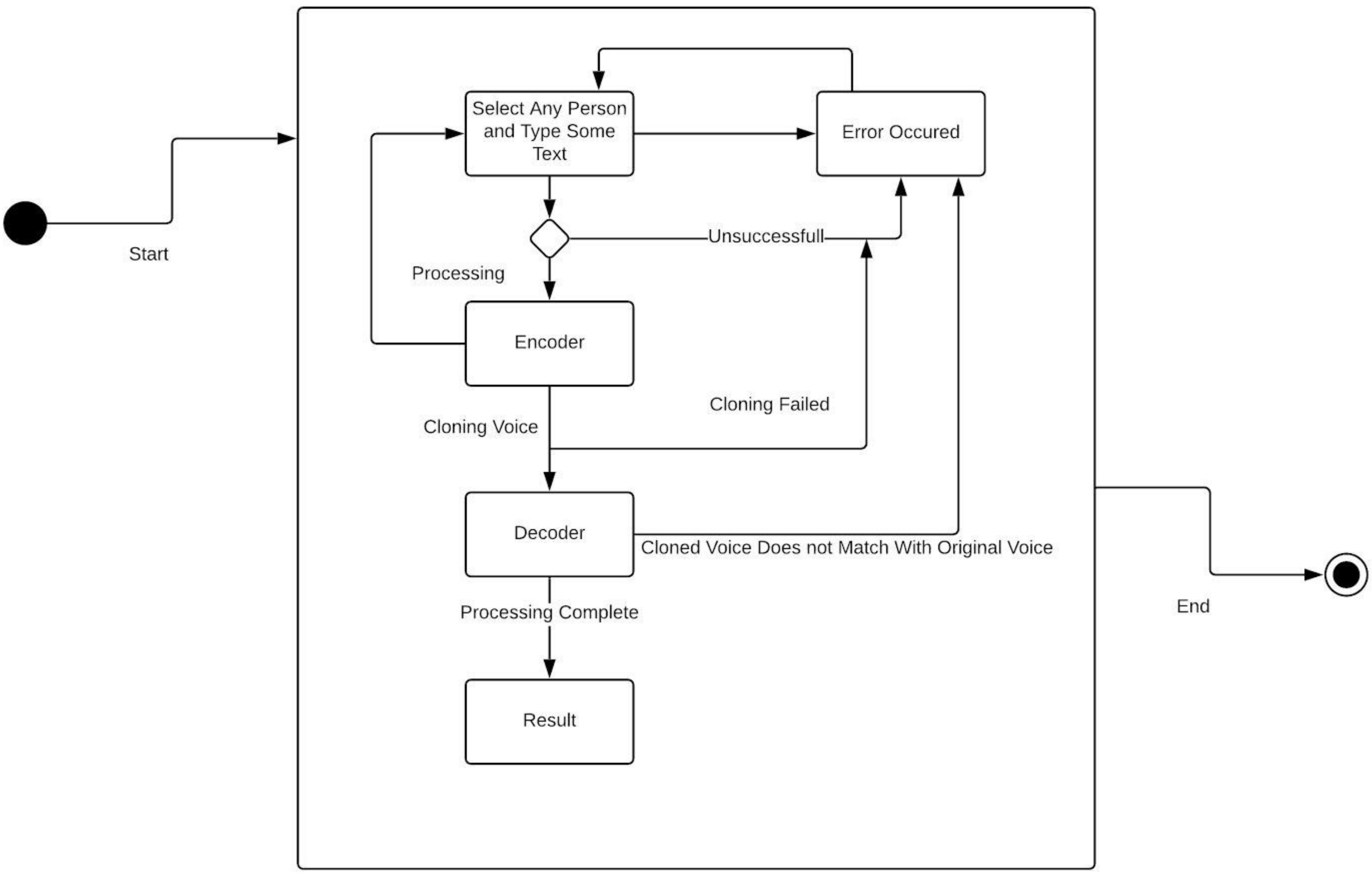
Class Diagram



Sequence Diagram



State Chart Diagram



Project outcome and Result Analysis

- The final project will be capable of converting and storing a digital copy of a person's natural voice.
- First step transforms the text into time-aligned features, such as mel spectrogram, or F0 frequencies and other linguistic features;
- Second step converts the time-aligned features into audio.
- We primarily used Tacotron 2 and WaveGlow models to achieve the output.
- Table 1 and Table 2 compare the training performance of the modified Tacotron 2 and WaveGlow models with mixed precision and FP32.

Number of GPUs	Mixed Precision mels/sec	FP32 mels/sec	Speed-up with Mixed Precision	Multi-GPU Weak Scaling with Mixed Precision	Multi-GPU Weak Scaling with FP32
1	20,992	12,933	1.62	1.00	1.00
4	74,989	46,115	1.63	3.57	3.57
8	140,060	88,719	1.58	6.67	6.86

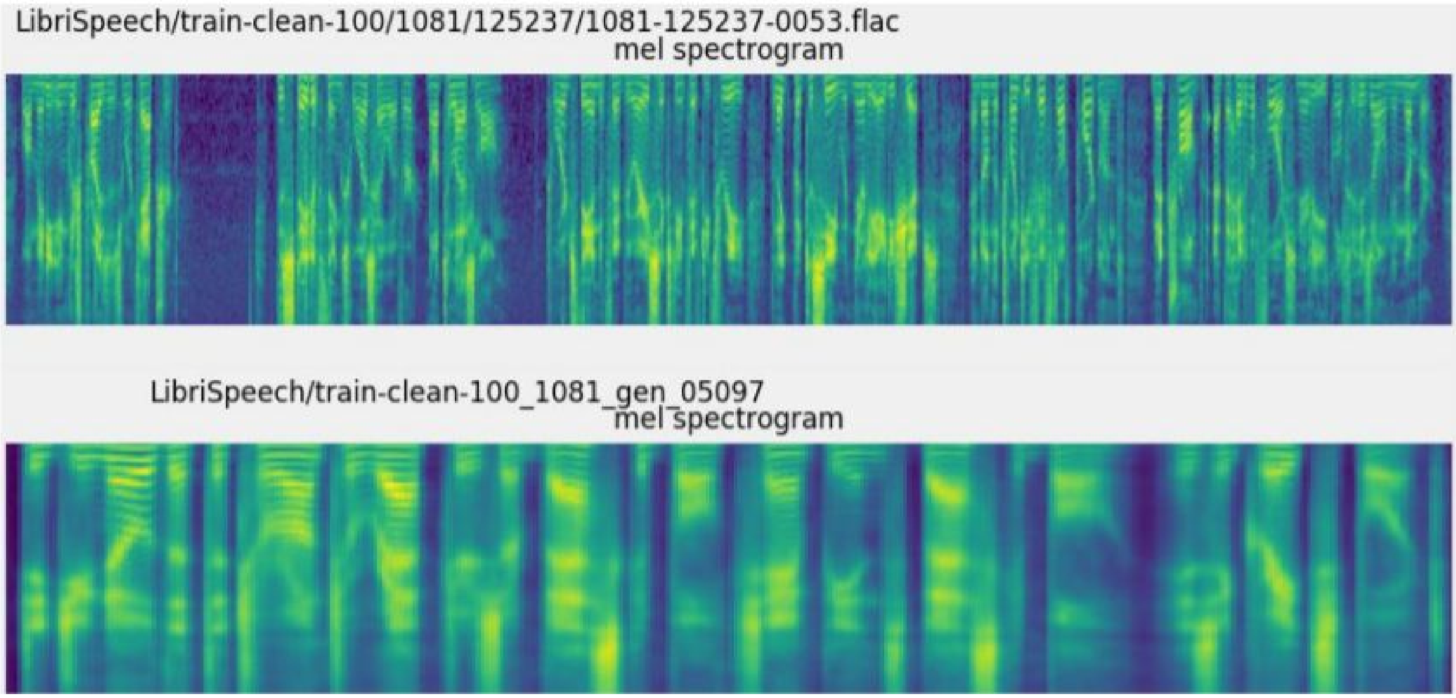
Table 1: Training performance results for modified Tacotron 2 model

Number of GPUs	Mixed Precision samples/sec	FP32 samples/sec	Speed-up with Mixed Precision	Multi-GPU Weak Scaling with Mixed Precision	Multi-GPU Weak Scaling with FP32
1	81,503	36,671	2.22	1.00	1.00
4	275,803	124,504	2.22	3.38	3.40
8	583,887	264,903	2.20	7.16	7.22

Table 2: Training performance results for WaveGlow model

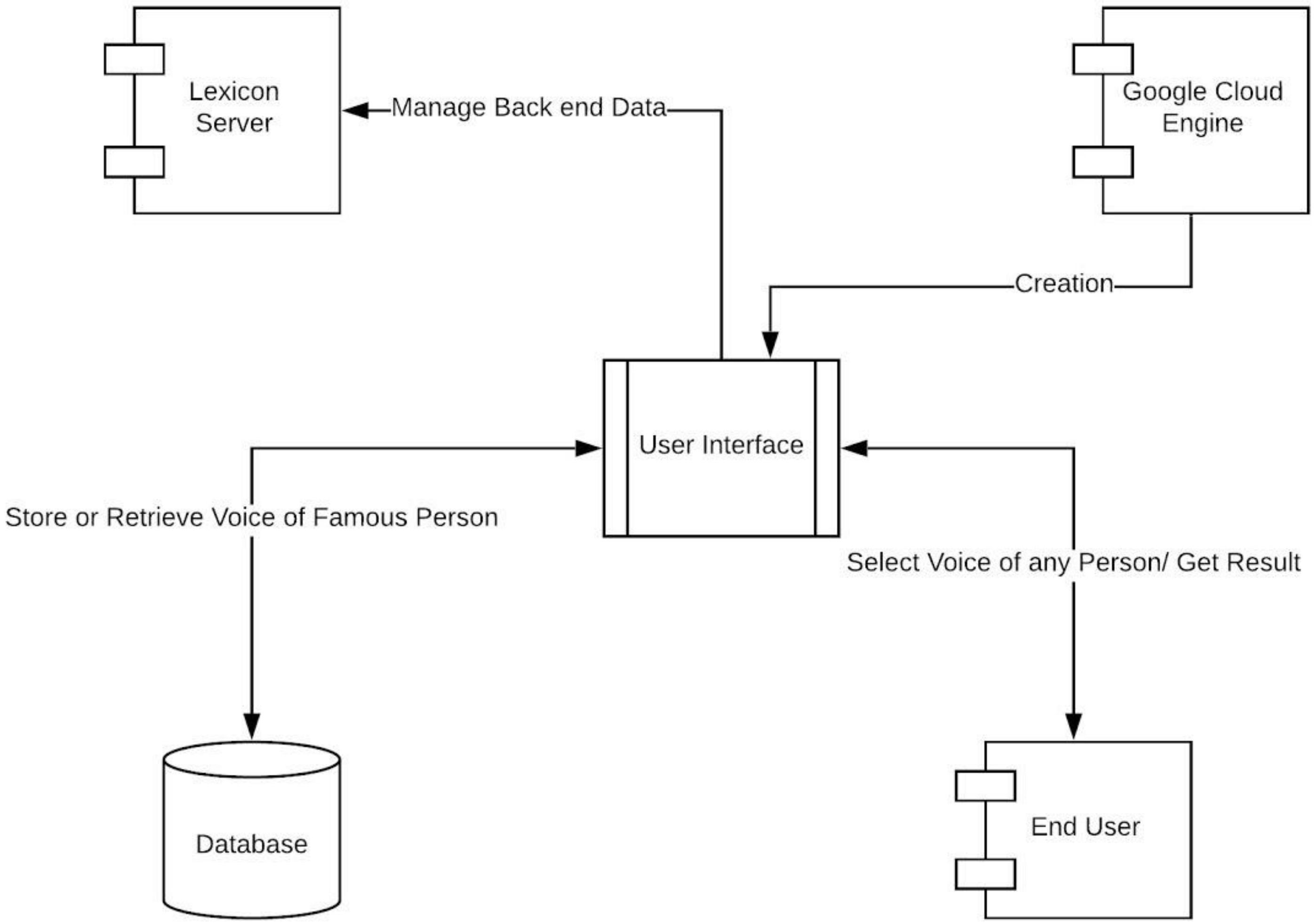
- As shown in Table 1 and 2, using Tensor Cores for mixed precision training achieves a substantial speedup and scales efficiently to 4/8 GPUs. Mixed precision training also maintains the same accuracy as single-precision training and allows bigger batch size.

- The waveform of the voice samples of original and deep faked audio would look like the image shown below:

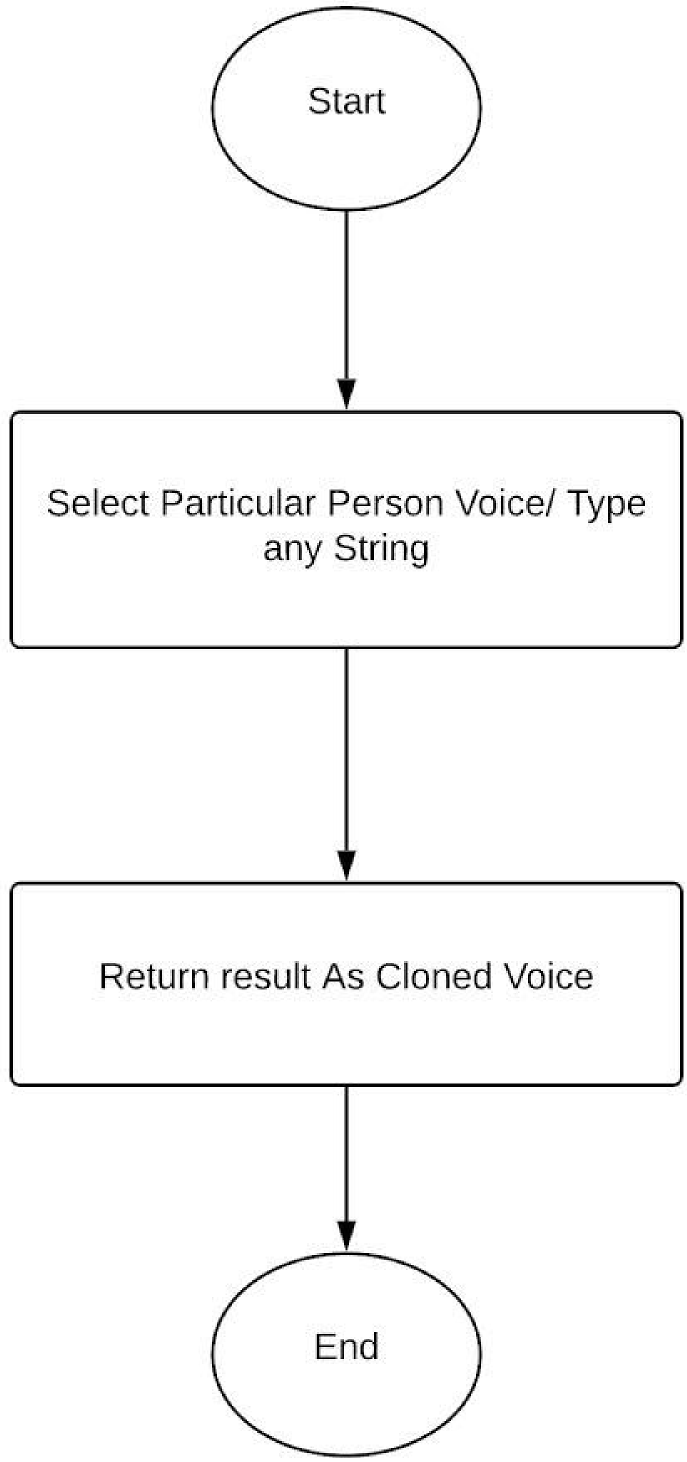


- Speech quality depends on model size and training set size; using Tensor Cores with automatic mixed precision makes it possible to train higher quality models in the same amount of time.

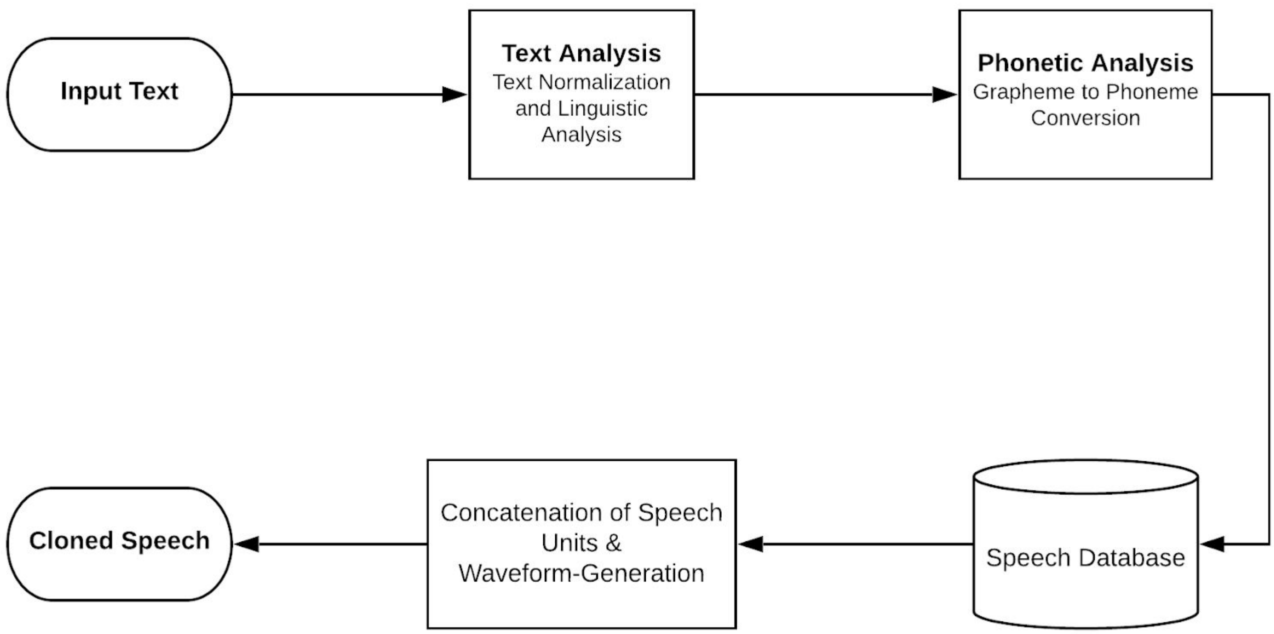
Component Design



Interface Design



Context Diagram



Tier Architecture