# Multi-Speaker Neural Voice Cloning in Nepali Language

**Aayush Man Shrestha (THA077BCT003)**
**Aditya Bajracharya (THA077BCT006)**
**Projan Shakya (THA077BCT037)**

**Under Supervision Of:**
**Er. Dinesh Baniya Kshatri**
**Lecturer**

Department of Electronics and Computer Engineering
Institute of Engineering, Thapathali Campus

March 2024

# Presentation Outline

- Motivation
- Introduction
- Problem Statement and Objectives
- Scope of Project
- Project Applications
- Methodology
- Results
- Discussion and Analysis
- Future Enhancements
- Conclusion
- Project Timeline
- References

# Motivation


Apps for English Voice Cloning

- Voice cloning is a fast-growing technology customizing digital voices for various needs.
- Nepali language awaits voice cloning innovations for greater inclusiveness.

# Introduction

- Process of using AI to replicate a person’s voice.
- Capture various nuances like pronunciation, nuances, rhythm and speech style.
- Takes input text and target audio from the user.
- Synthesizes speech for the input text in target speaker’s voice.

# Problem Statement and Objectives

- Problem Statement
  - Absence of voice cloning in Nepali language
- Objectives
  - To collect and organize datasets in Nepali language for voice cloning
  - To train a multi-speaker generative model for voice cloning with the collected datasets

# Scope of Project

## Project Capabilities:

- Clones voice of target person in Nepali language
- Generates cloned speech for provided Devanagari text
- Requires very limited amount of data to clone voice

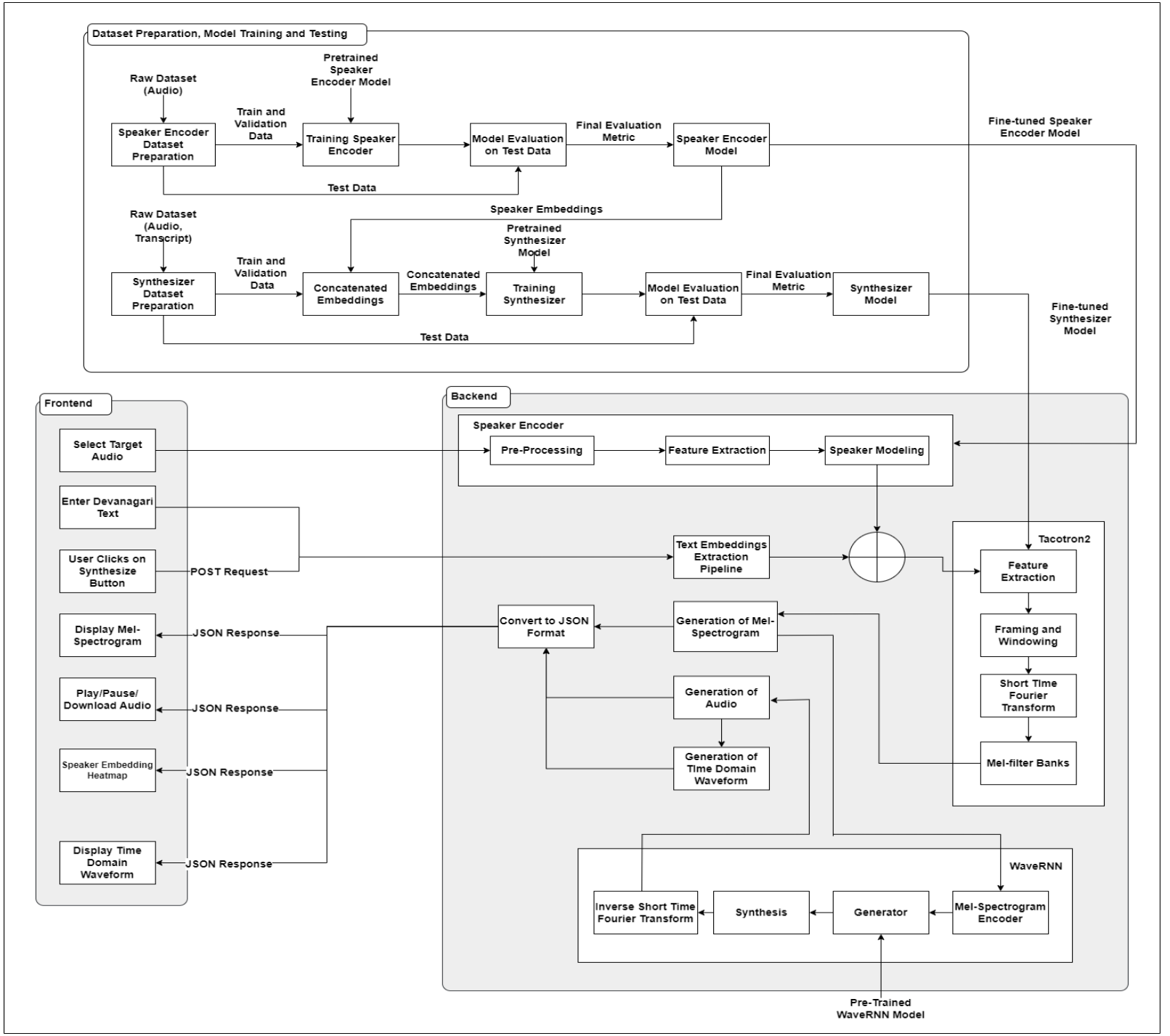
## Project Limitations:

- Inaccurate for words with complex pronunciation
- Cannot replicate complex speech patterns
- Requires high quality target audio

# Project Applications

- Language Preservation and Documentation
    - Capture voice of native speakers with unique dialects and accents
    - Replicate voices of notable figures
- Entertainment Industry
    - Dub Nepali movies and TV shows
- Content Creation
    - Podcasting, YouTube videos

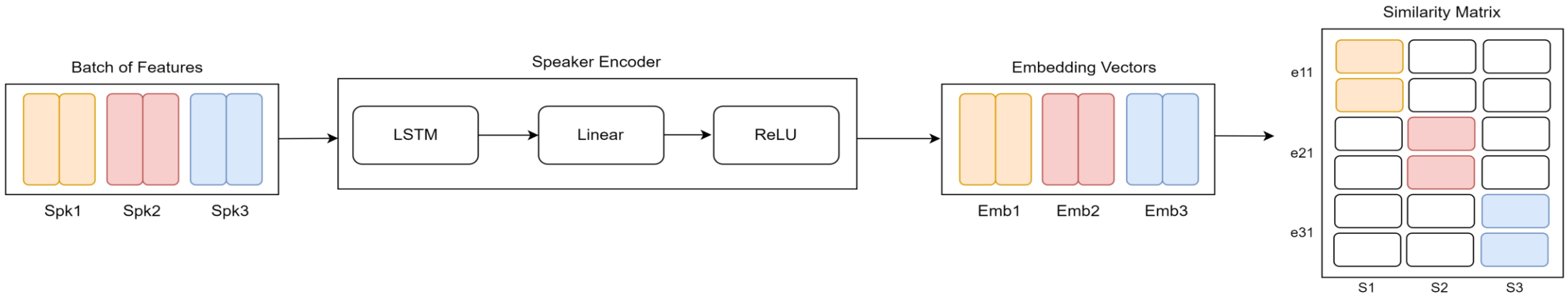
# Methodology – [1] (System Block Diagram)



# Methodology – [2] (Working Principle)

- User types the input text: Preeti or Unicode
- User provides target audio: Record, Select speaker or Upload
- Speaker encoder generates speaker embeddings
- Tacotron2 takes text and speaker embeddings
- Generate text embeddings and concatenate with speaker embeddings
- Generate mel-spectrogram and alignment graph
- Generate audio using WaveRNN vocoder

# Methodology – [3] (Speaker Encoder)

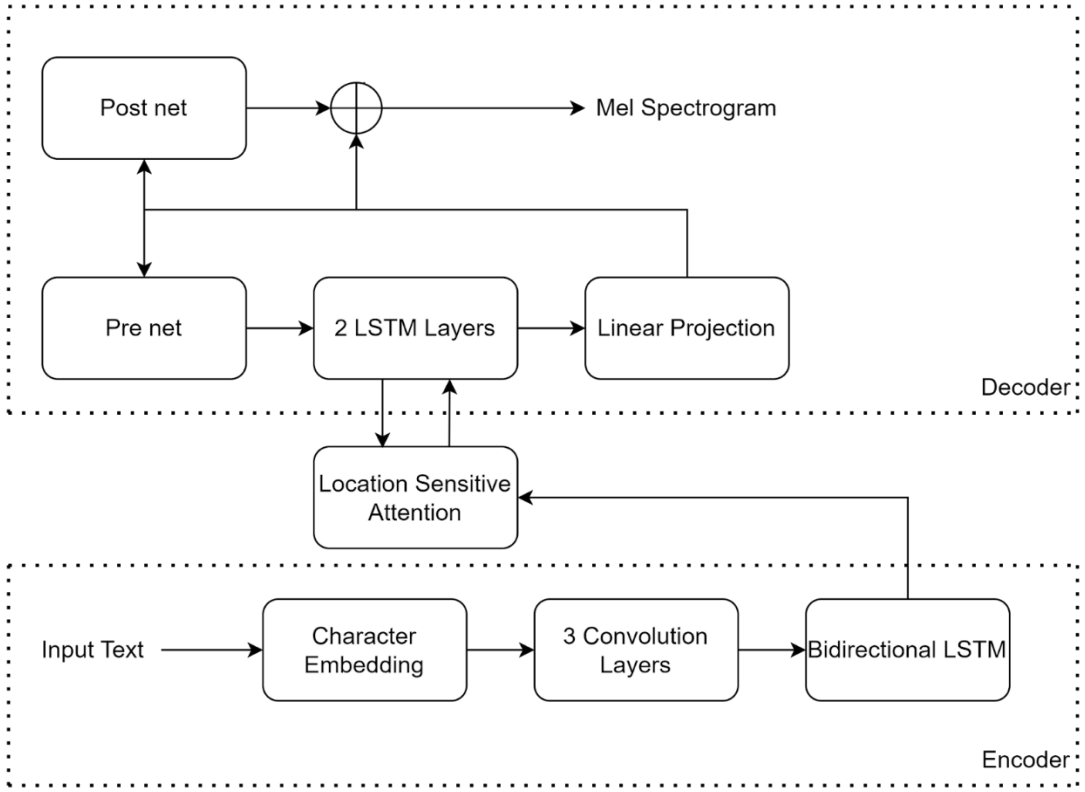



- 3 LSTM each containing 256 units.
- Extracts speaker-specific features: Pitch, Prosody and Speech rate
- One fully connected layer (256 units) with ReLU activation
- Maps the output into the desired embedding space.

# Methodology – [4] (Similarity Matrix)

- Quantifies the similarity relationships between embeddings of utterances in a batch.
- Cosine similarities between embeddings and centroids representing speakers.
- Two types of centroids are computed:
  - Inclusive centroid: consider all embeddings including the current utterance
  - Exclusive centroid: exclude the current utterance's embedding
- Compares utterances between same speaker and different speakers.
- Ensure embeddings capture speaker characteristics while remaining distinct.

# Methodology – [5] (Tacotron2 Architecture)



# Methodology – [6] (Tacotron2 - Encoder)

- Character Embedding:
  - Converts input text into embedding
  - Fixed dimension of 512
- Convolution Layers:
  - Three 1D convolution layers
  - Each with 512 filters of size 5x1
  - Stride size of 1, dilation of 1
  - ReLU activation
- Bidirectional LSTM:
  - Single Bi-LSTM layer 512 units
  - Processes data in both forward and backward directions
  - Captures context of each character
  - Dropout rate of 0.1 for regularization

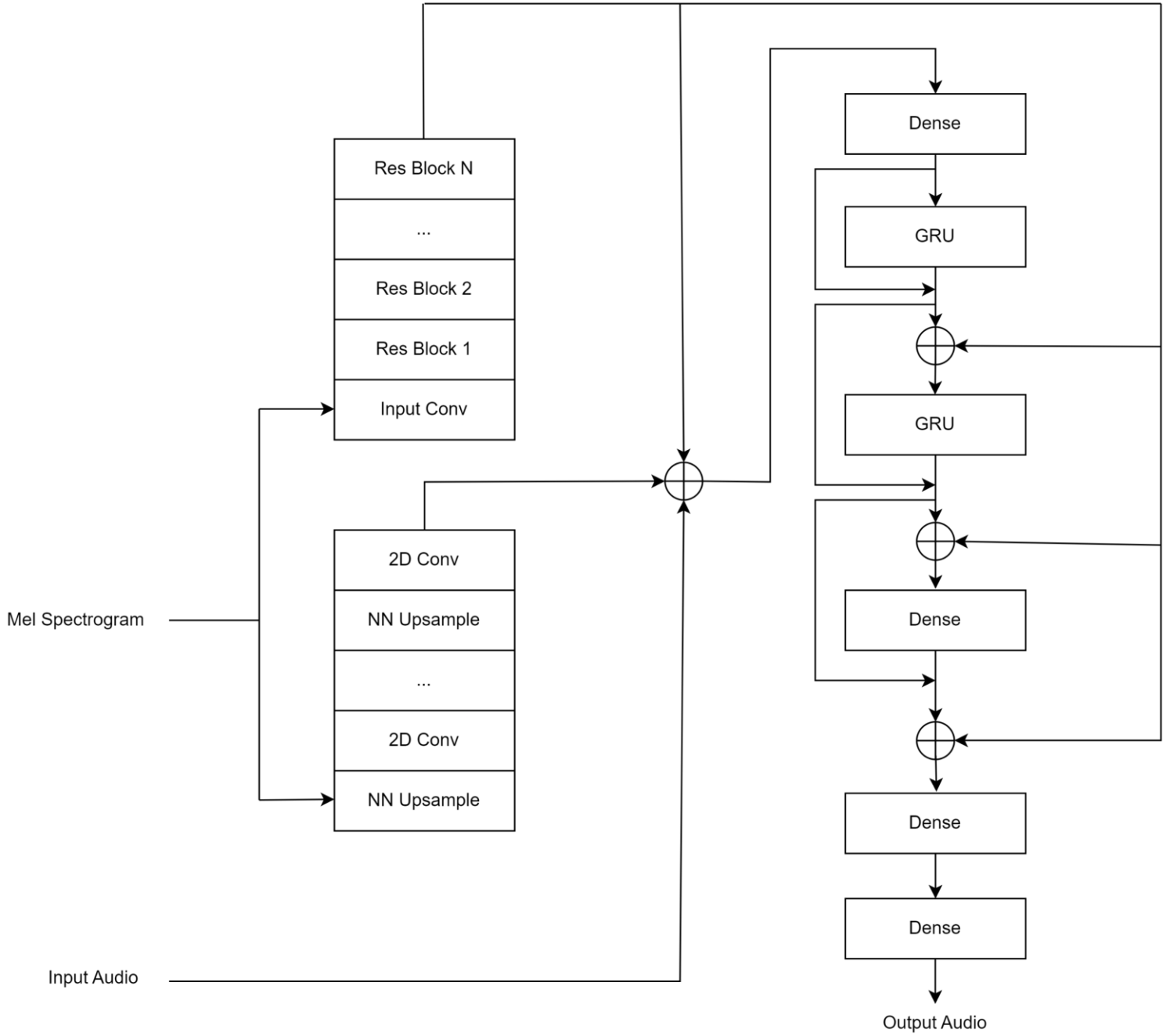
# Methodology – [7] (Tacotron2 - Attention)

- Summarizes encoded sequence for decoder
- Location layer for positional information:
  - Composed of 32 1-D convolution filters (length: 31)
  - Captures context of prior attention weights
- Attention computation:
  - Dot product of query and values
  - SoftMax activation

# Methodology – [8] (Tacotron2 - Decoder)

- Decoder Prenet
  - Two fully connected layers with ReLU activations
  - Operates on attention mechanism output
- Unidirectional LSTM
  - Two uni-directional LSTM layers with 1024 units
  - Mel spectrogram frame prediction and autoregressive generation
- Decoder Postnet
  - Enhances spectrogram reconstruction
  - Predicts residual for refinement

# Methodology – [9] (WaveRNN Architecture)



# Methodology – [10] (WaveRNN Architecture)

## UpSampling Network

- Enhances temporal resolution of mel spectrograms.
- Consists of 3 convolutional layers (128 units each) with kernel size 1 and padding 2.
- First 2 layers upsample by factor 5, third layer by factor 8.
- Employs linear layers and ReLU activation.

## Resnet Blocks

- Extracts features from mel spectrograms.
- Utilizes 10 ResNet blocks with 2 convolution layers (128 units, kernel size 1) each.
- Incorporates skip connections for efficient gradient flow.
- Enables deeper architectures without performance degradation.

# Methodology – [11] (WavRNN Architecture)

- Incorporates GRU layers and dense layers.
- Concatenates audio signal and extracted features.
- Features 4 dense layers (512 unite each) for dimensional transformation and feature mapping.
- Employs 2 GRU layers (512 units each) processing sequential information.
- Concatenates GRU outputs with auxiliary features for improved discriminative learning.

# Data Exploration – [1]
# (Nepali Datasets for Voice Cloning)

**Speaker Encoder Dataset**

- OpenSLR Datasets
  - SLR54
- Self-Collected Data
  - Audiobooks
  - Self-Recorded
  - YouTube: Interviews, Podcasts


**Synthesizer Dataset**

- OpenSLR Datasets
  - SLR43
  - SLR143
- Self-Collected Data
  - Audiobooks
  - Self-Recorded

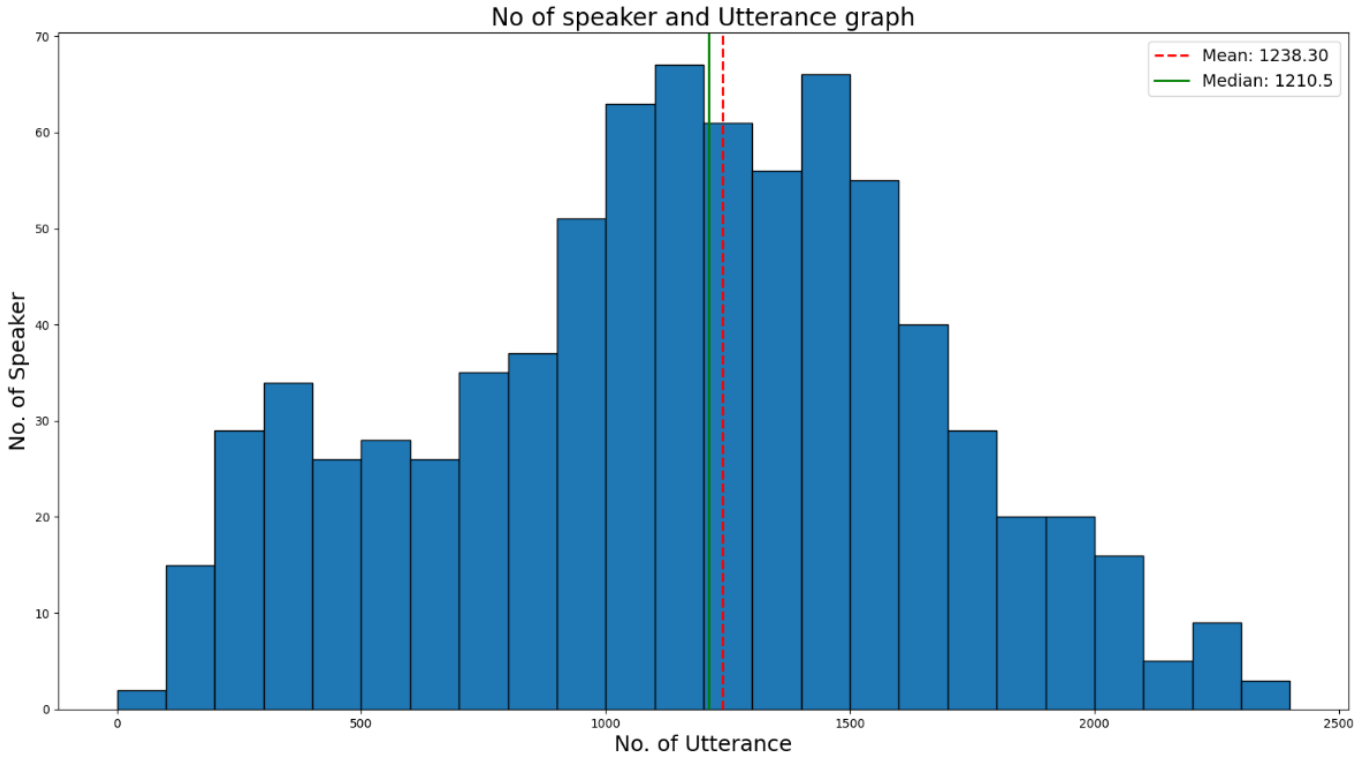
# Data Exploration – [2]
# (Glimpse of Speakers)

| Audiobooks | Author | Speaker(s) |
|---|---|---|
| भैरव अर्यालका हास्यव्यङ्ग्य | Bhairab Aryal | Ananda Nepal |
| पर्दाफास | Salina Thapa | NBSA Speaker |
| हिटलर र यहुदी | B.P Koirala | Neu Narrations |
| सुम्निमा | B.P Koirala | Neu Narrations |
| श्वेतभैरवी | B.P Koirala | Cleartech |

# Data Exploration – [3]
# (Speaker Encoder Datasets for Voice Cloning)

| Audio Statistics | Self Collected | OpenSLR54 |
|---|---|---|
| **Total Speakers** | 306 | 527 |
| **Male Speakers** | 165 | 316 |
| **Female Speakers** | 141 | 211 |
| **Average Audio per Speaker** | 14 min 36 sec | 17 min 36 sec |
| **Total Duration** | 70 hours | 165 hours |
| **Longest Audio of Single Speaker** | 58 min 51 sec | 41 min 48 sec |
| **Shortest Audio of Single Speaker** | 2 min 05 sec | 49 sec |
| **Total Utterance** | 338469 | ~157000 |

# Data Exploration – [4]
# (Speaker Encoder Datasets for Voice Cloning)

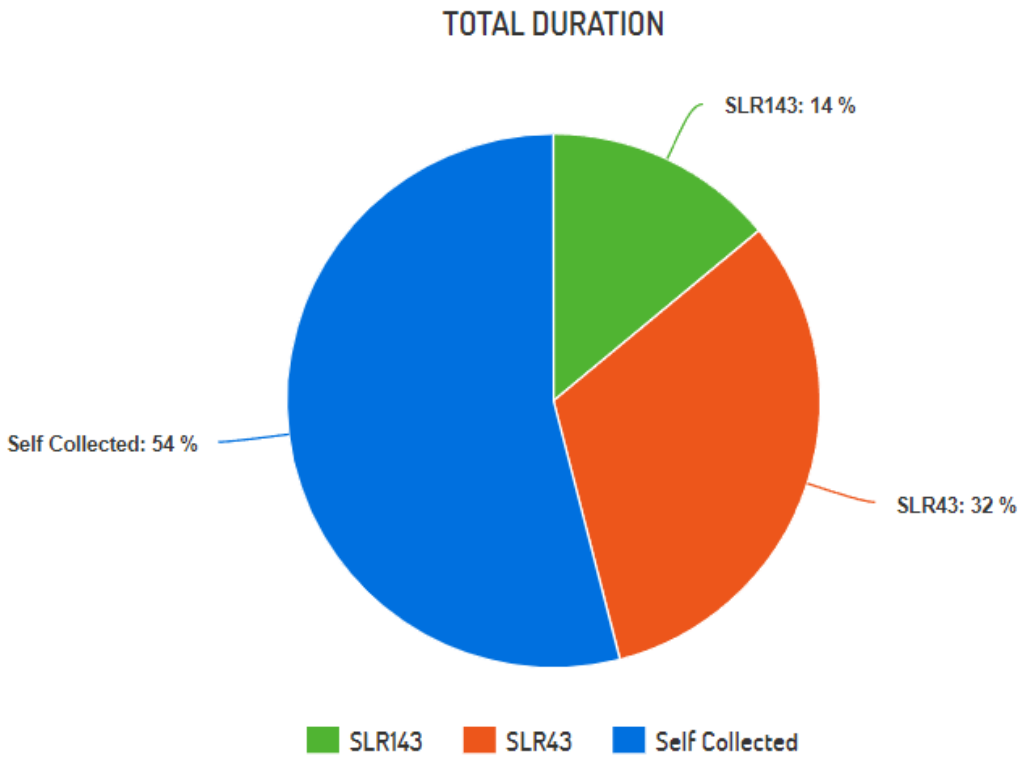


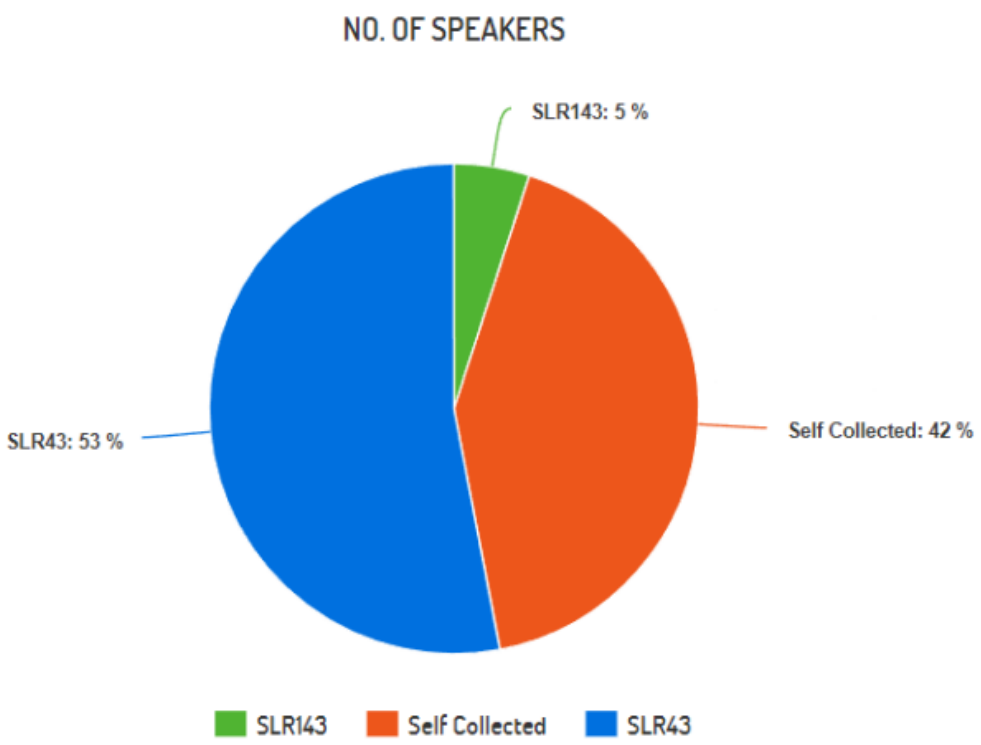

- Has 1238.30 utterance per speaker
- Has median of 1210.5 utterance

# Data Exploration – [5] (Synthesizer Dataset Summary)

| S. N. | Dataset Description | No. of Speakers | Total Duration (hrs) | Average Duration (sec) | Sampling Rate (kHz) | Audio Format |
|---|---|---|---|---|---|---|
| 1 | SLR43 | 19 | 2.80 | 5 | 48 | wav |
| 2 | SLR143 | 2 | 1.20 | 6 | 22.05 | wav |
| 3 | Self-collected Dataset | 15 | 4.67 | 5 | 22.05 | wav |

# Data Exploration – [6] (Dataset Summary)



# Data Exploration – [7]
# (Self-collected Data Overview)

| S.N. | Source | Total Duration (hrs) | Number of Speakers | Number of Datasets | Audio Format |
|---|---|---|---|---|---|
| 1 | Audio Books | 3.5 | 5 | 2712 | wav |
| 2 | Voice Recordings | 1.17 | 10 | 595 | wav |

# Data Exploration – [8] (Metadata of Dataset)

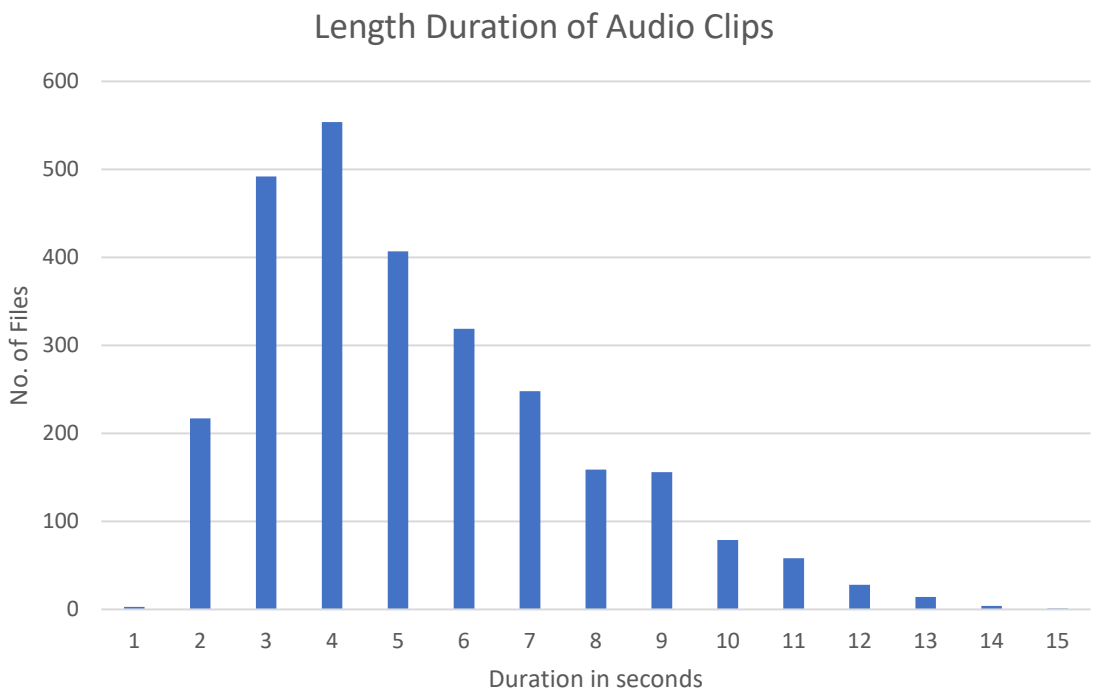
| S.N. | Dataset | Male Distribution (%) | Female Distribution (%) | Age Group | Children's Voice |
|---|---|---|---|---|---|
| 1 | SLR43 | 0 | 100 | 20-40 | No |
| 2 | SLR143 | 20 | 80 | 20-40 | No |
| 3 | Self-Collected | 75 | 25 | 20-40 | No |

# Data Exploration – [9] (Gathered Dataset Statistics)

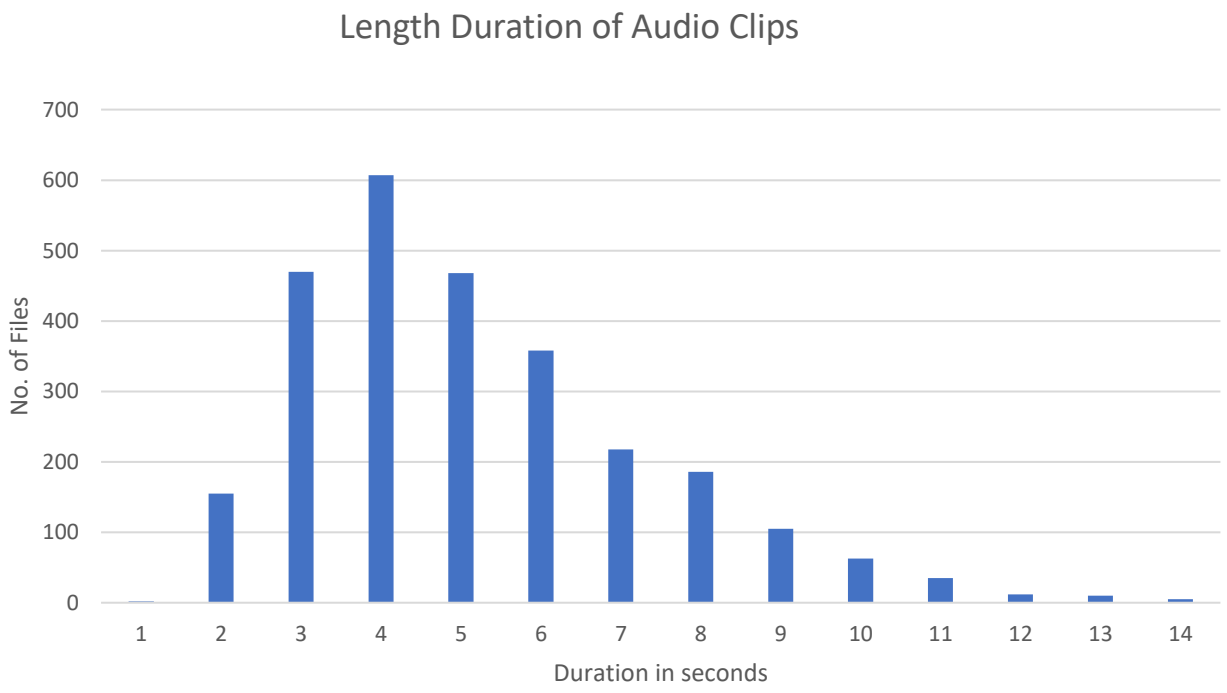
| | |
|---|---|
| **Total Clips** | 6046 |
| **Total Words** | 64,232 |
| **Total Characters** | 407,371 |
| **Total Duration** | 8.67 hours |
| **Mean Clip Duration** | 5.162 seconds |
| **Minimum Clip Duration** | 1.404 seconds |
| **Maximum Clip Duration** | 14.542 seconds |
| **Mean Words per Clip** | 10.63 |
| **Distinct Words** | 79,375 |

# Data Exploration – [10] (Dataset Comparison)

## OpenSLR Datasets

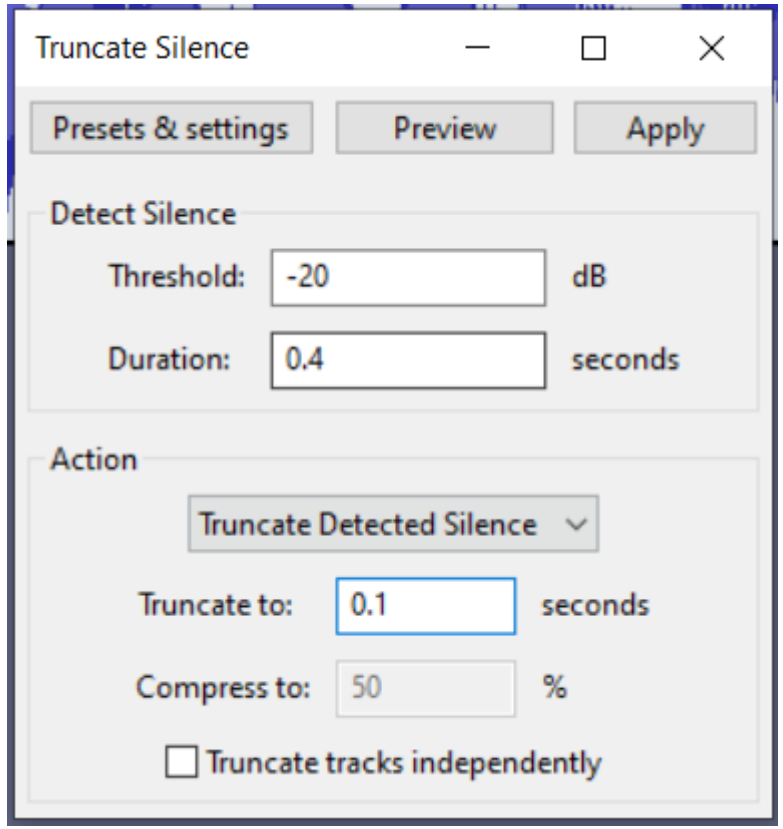



## Self-collected Datasets



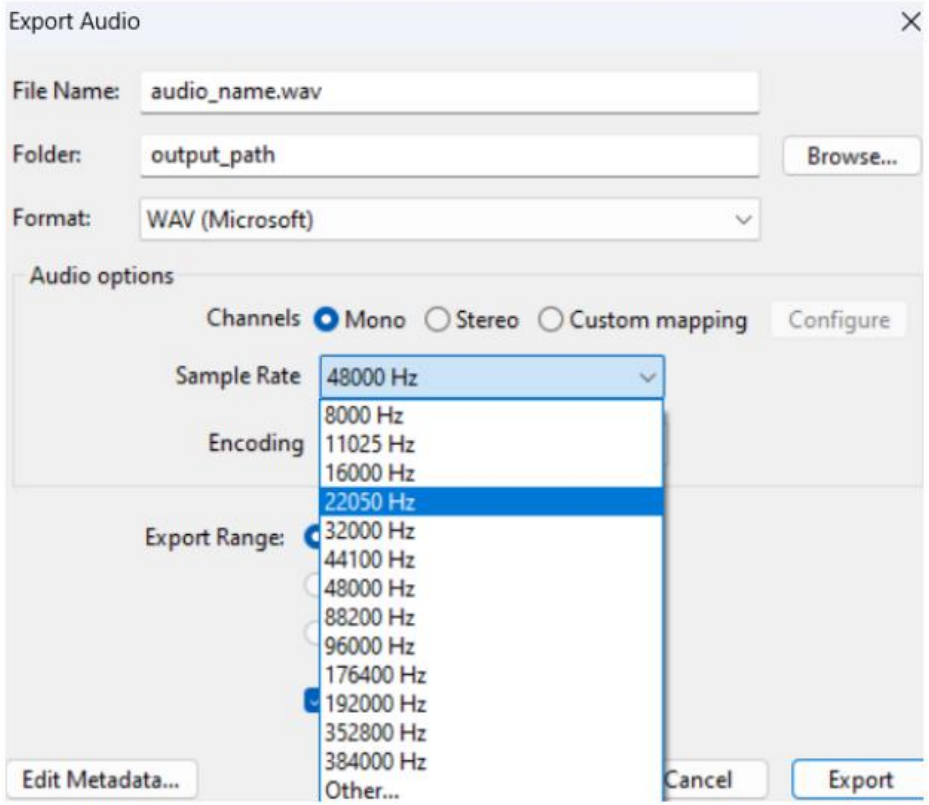
# Data Pre-processing – [1] (Speaker Encoder Audio Pre-processing)

- Resampling
  - Resampled to 16 KHz for Speaker Encoder
- Clipping
  - Clipped to 1.6 sec of audio
  - 50% overlap
- Silence Removal
  - Silence more than 0.4 sec truncated to 0.1 sec



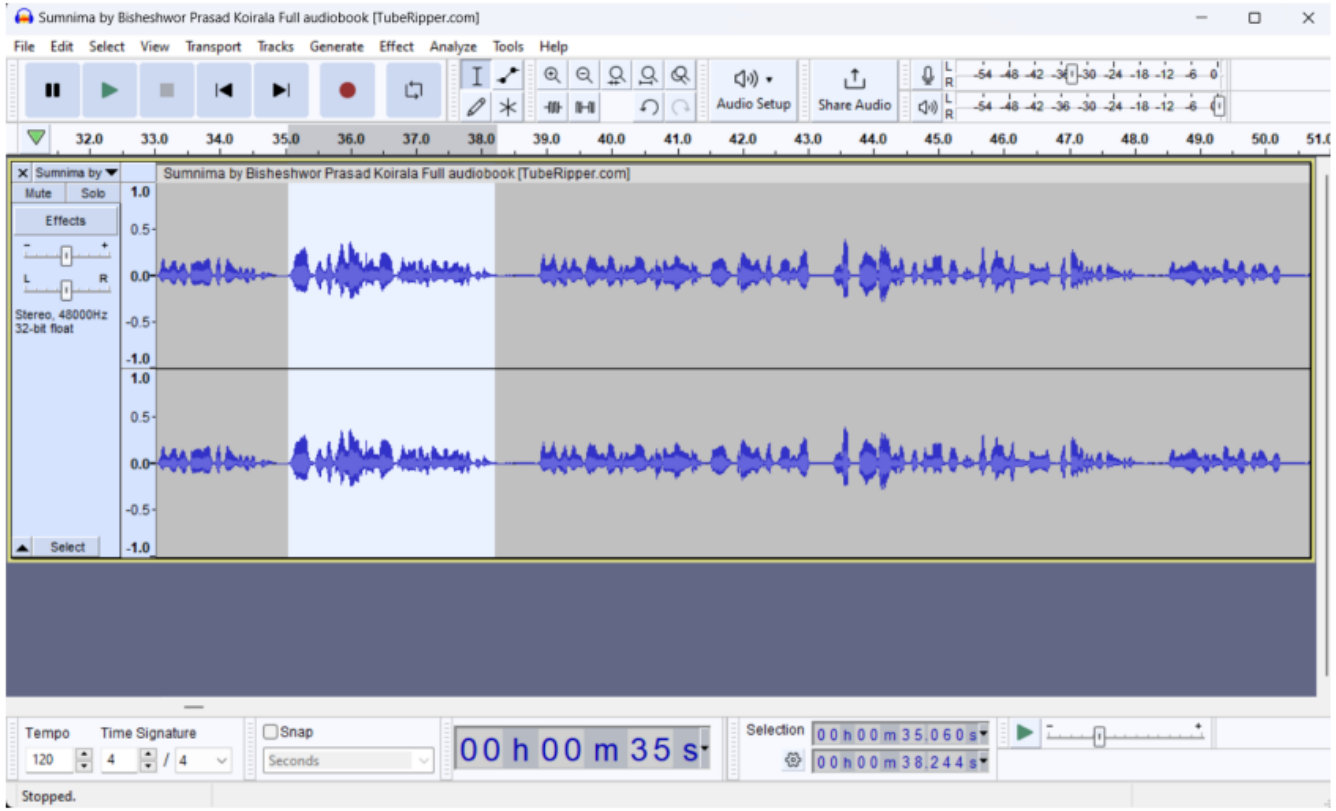
# Data Pre-processing – [2] (Synthesizer Audio Pre-processing)

- Resampling
  - Converting sampling frequency from one to another
  - Done to ensure uniformity and no loss of information
  - Sampled to 22.05 KHz
- Normalization
  - Making the amplitude consistent over the audio signal
  - Made to reduce volume variations



# Data Pre-processing – [3] (Synthesizer Audio Pre-processing)

- Noise Reduction and Filtering
  - Noisy signals clipped
  - Passed through LC filter to remove unwanted frequencies



# Data Pre-processing – [4]
# (Text Pre-processing)

| S. N. | Raw Text | Pre-Processed Text | Remarks |
|---|---|---|---|
| 1. | भर्खर ७ प्याग हुँदैछ, १५ प्याग खाएपछि मात्र मलाई थोरै थोरै लाग्छ । | भर्खर सात प्याग हुँदैछ, पन्ध्र प्याग खाएपछि मात्र मलाई थोरै थोरै लाग्छ । | Pre-processing numerals |
| 2. | बरु अब त अझ बढी लजालु भएकी थिएँ; साथीहरु पनि कम थिए । | बरु अब त अझ बढी लजालु भएकी थिएँ। | Breaking down long texts |
| | | साथीहरु पनि कम थिए । | |
| 3. | त्यो चिच्याहट सुनेर क्यालीगुला आनन्दित हुन्थ्यो | त्यो चिच्याहट सुनेर क्यालीगुला आनन्दित हुन्थ्यो । | Adding stop tokens |

# Instrumentation – [1] (Hardware Requirements)

Cloud Computing Resources used:

**Google Colab**

- GPU: Tesla T4
- GPU Memory: 16 GB
- RAM: 12 GB
- Session: 6 hours

**Kaggle**

- GPU: Tesla T4 x2, P100 (30 hours a week or 9 hours a session)
- TPU: VM v3-2 (20 hours a week or 3 hours a session)
- RAM: 29 GB
- Session: 12 hours

# Instrumentation – [2] (Software Requirements)

| Purpose | Tools Used |
|---|---|
| Text Editor | VS Code |
| Programming Language | Python |
| Audio Processing | Librosa, Audacity |
| Framework | PyTorch |

| Purpose | Tools Used |
|---|---|
| Data Visualization | Matplotlib, TensorboardX |
| Frontend | HTML, CSS, JavaScript |
| Backend | Uvicorn, FastAPI |
| Testing and Debugging | Postman, Chrome Browser Console |

# Evaluation – [1]
# (Qualitative Evaluation - MOS)

- Stands for Mean Opinion Score
- Value Range: 1 to 5
- Formula for MOS:

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

Where, $R_n$ = MOS Rating of $n^{th}$ participant

N = Total number of participant

# Evaluation – [2] (MOS Scores)

| MOS Score | Speech Quality | Speech Similarity |
|---|---|---|
| 1 | Can’t understand | Definitely not same person |
| 2 | Some words unclear, pauses and pronunciation issue | Not sure if same person |
| 3 | Generally understandable and acceptable | Same person, but a bit different |
| 4 | Natural, clear, and understandable, good hearing | Highly likely same person, tone and intonation similar |
| 5 | Broadcasting level | Definitely same person, tone and speaking style match |

# Evaluation – [3] (Quantitative Evaluation)

- Correlation

$$SignalValidation = correlationf(abs(fft(x_1)), abs(fft(x_2)))$$

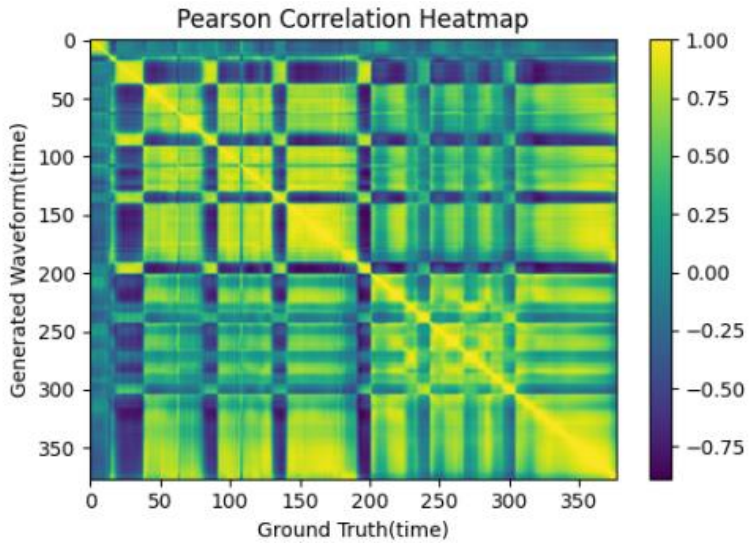
  - Value: Ranges from +1 to -1
    - Closer the value to +1, higher the similarity
    - Closer the value to -1, lower the similarity
  - Formula for Correlation:
    - $r = \frac{n \sum xy - \sum x \sum y}{\sqrt{[(n \sum x^2 - (\sum x)^2][(n \sum y^2 - (y)^2]}}$

Where, $r$ = Karl Pearson's correlation coefficient

$x = abs(fft(x_1))$

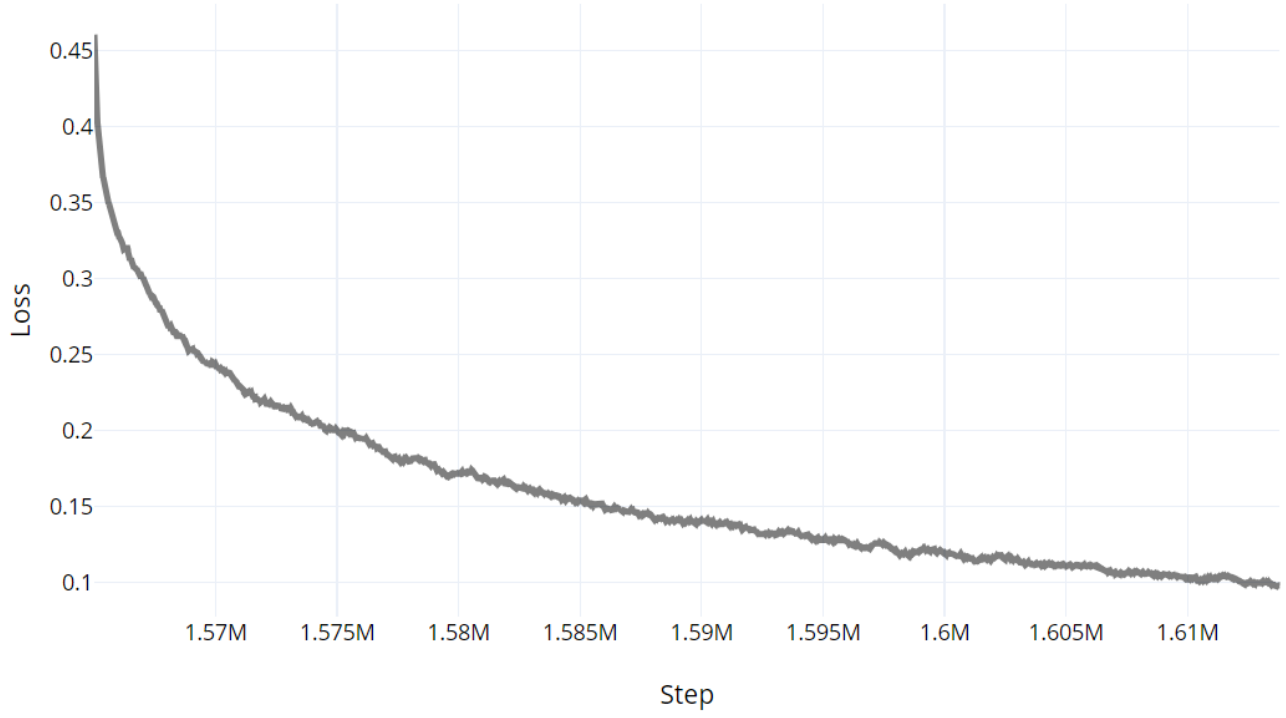
$y = abs(fft(x_2))$

$n$ = Number of data points



# Results – [1]
# Speaker Encoder Training

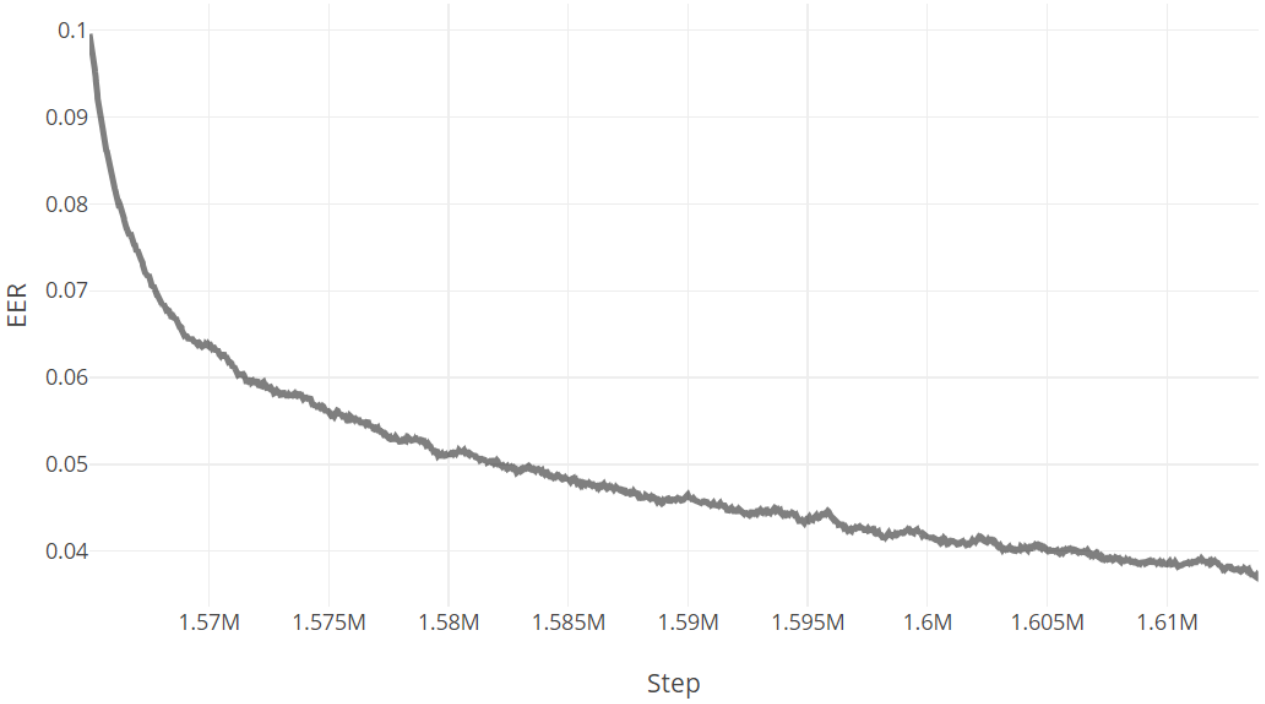
Loss Curve for Speaker Encoder




- Loss curve started from 1.57M due to pre-trained model
- Loss decreasing gradually

# Results – [2]
# Speaker Encoder Training

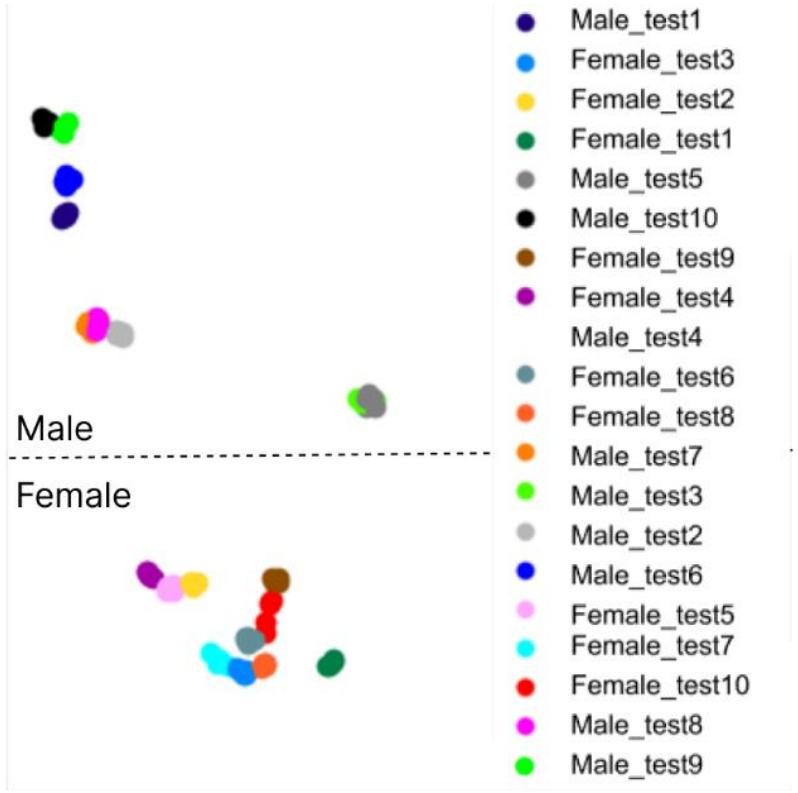
EER Curve for Speaker Encoder




- EER curve started from 1.57M due to pre-trained model
- EER also decreasing gradually

# Results – [3]
# Speaker Encoder Training

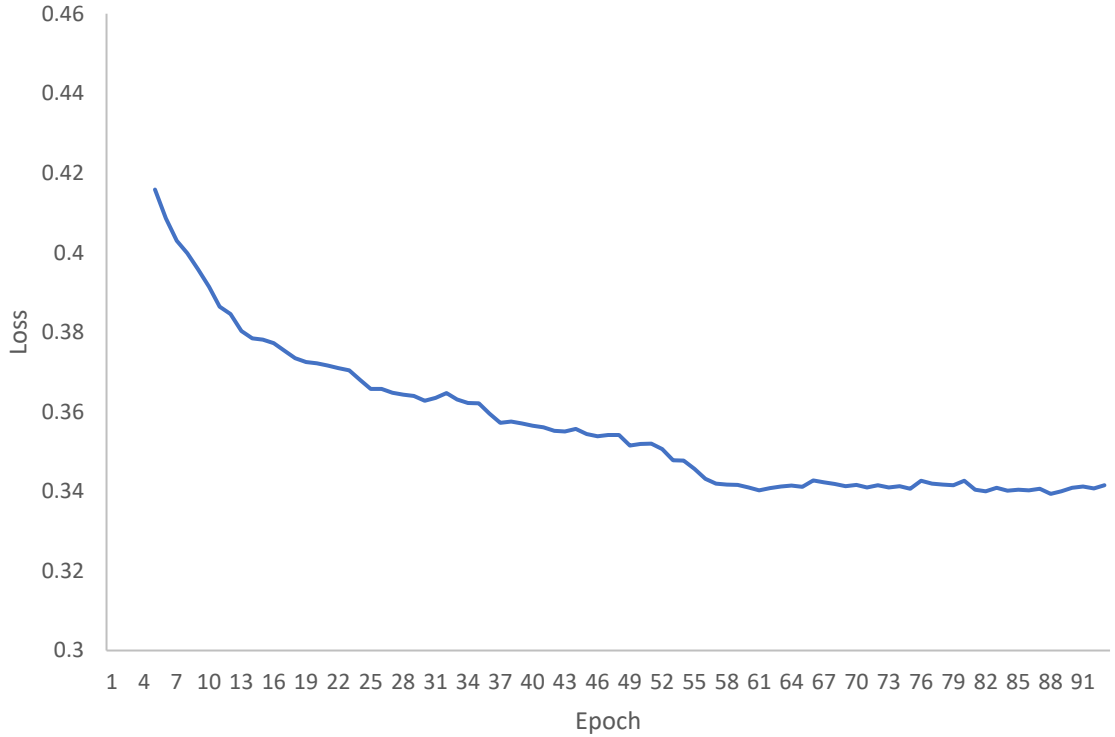
UMAP of Test Dataset




- Same speaker utterance are grouped
- Male and female are separated

# Results – [4] Synthesizer Training

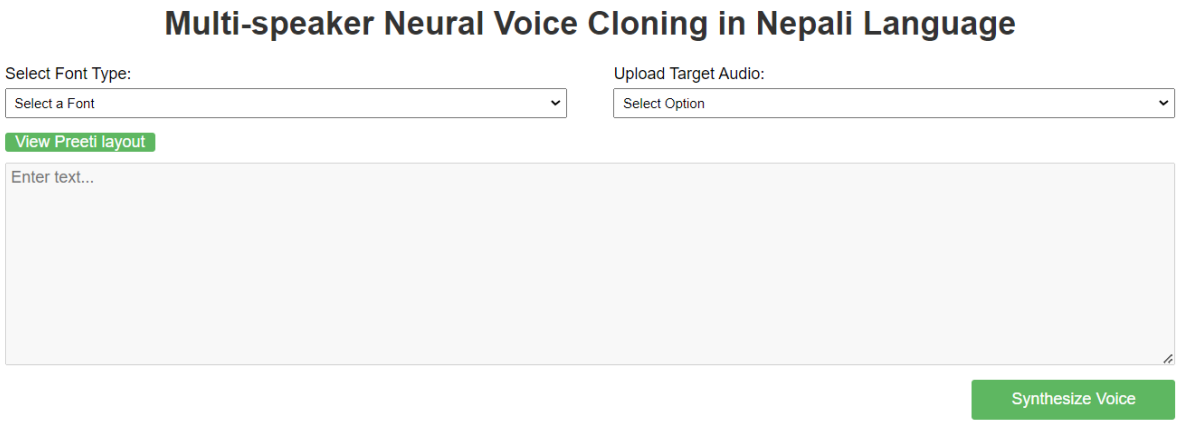
Loss Curve of Synthesizer

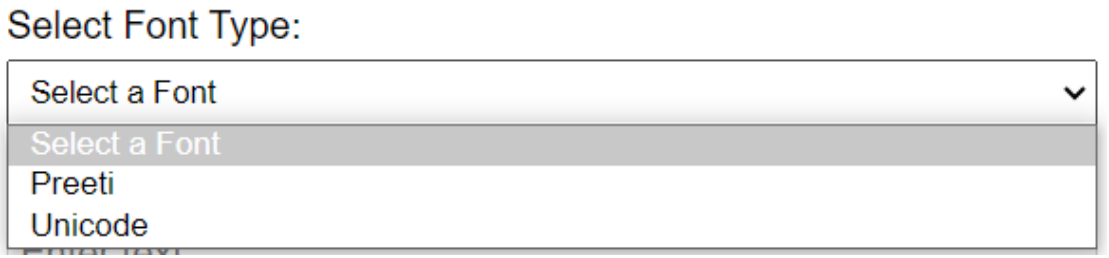



- Synthesizer trained till 92nd epoch
- Loss started from 0.42 and dropped till 0.34

# Results – [5]
# (Web Interface)

## Input Section

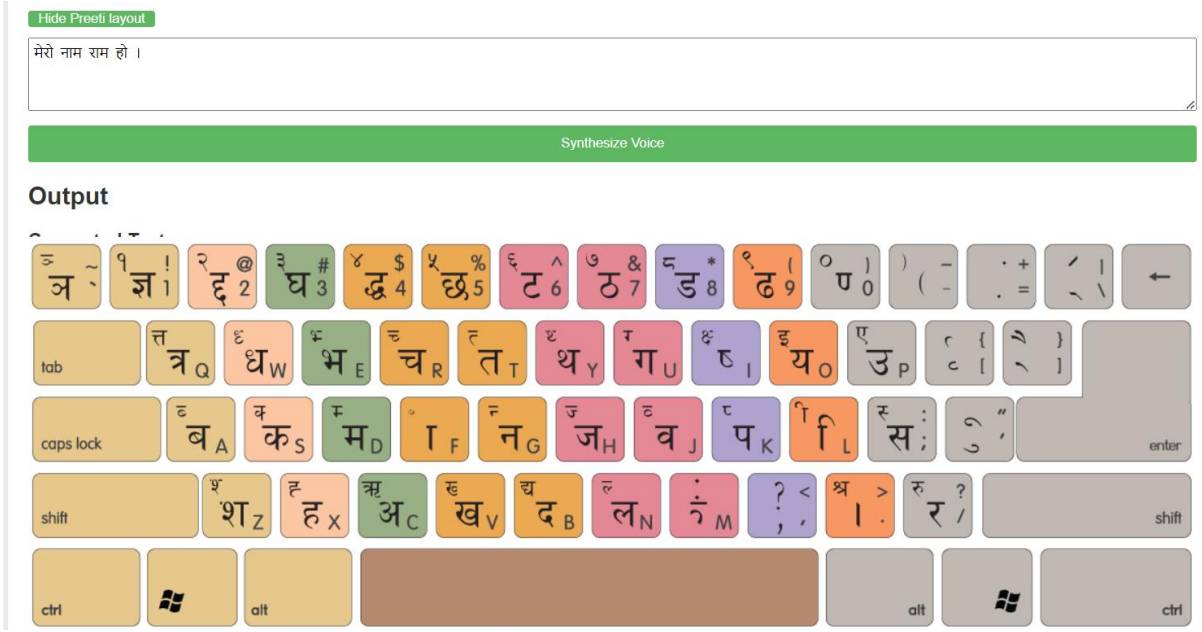



- Has two option to enter text
- Preeti for basic Devanagari font
- Unicode for Romanized English

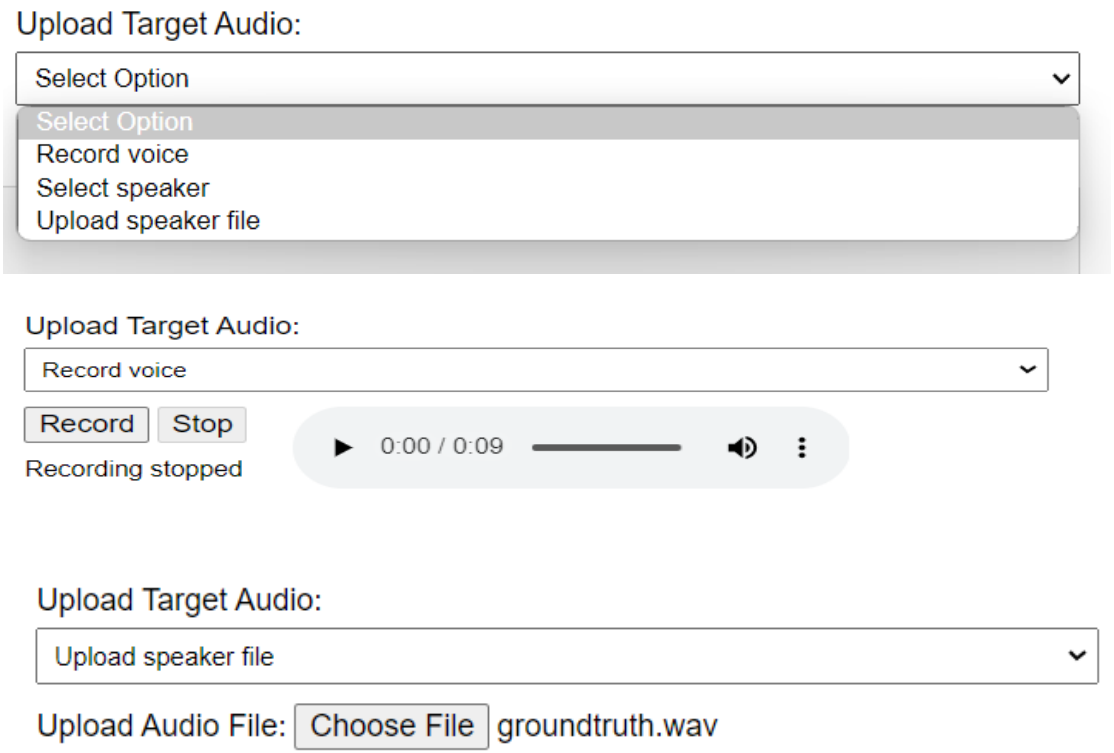
# Results – [6]
# (Web Interface)

## Font Selection



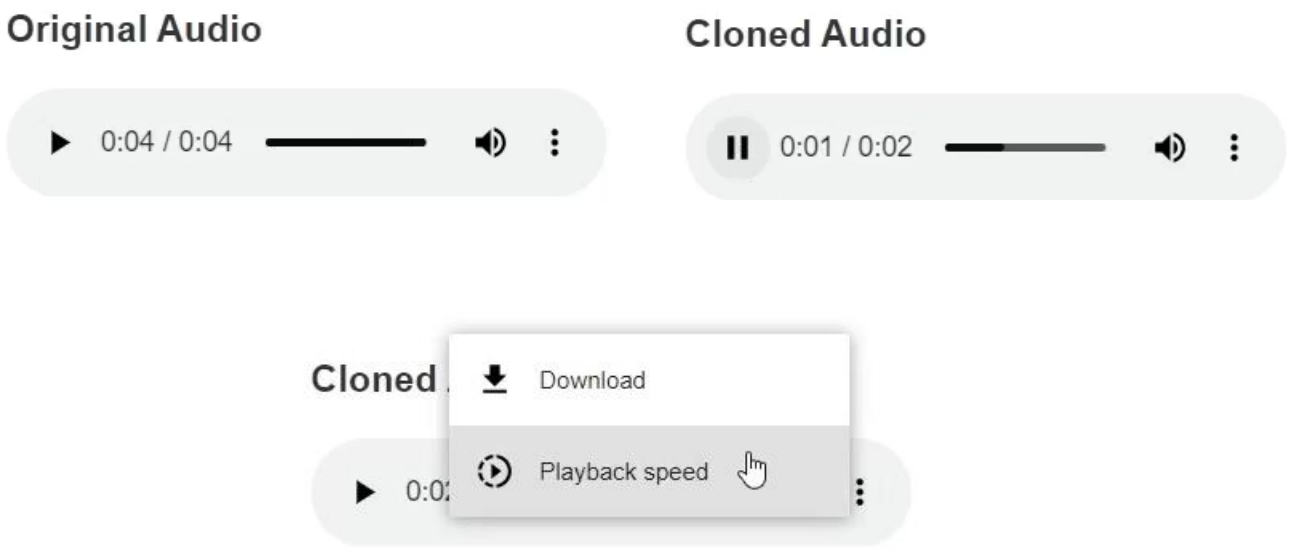

# Results – [7] (Web Interface)

## Target Audio Selection



# Results – [8] (Web Interface)

Audio Players

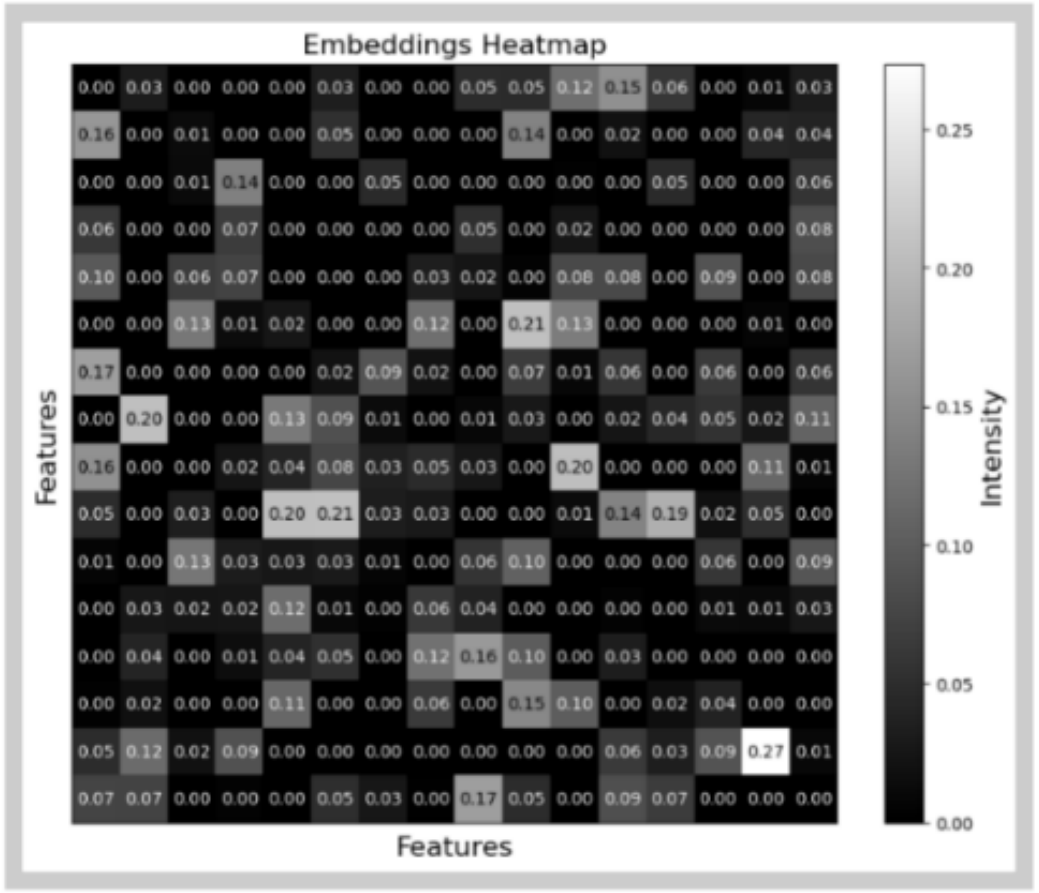



- Audio file player can play both original and cloned audio
- Can download the audio or change playback speed

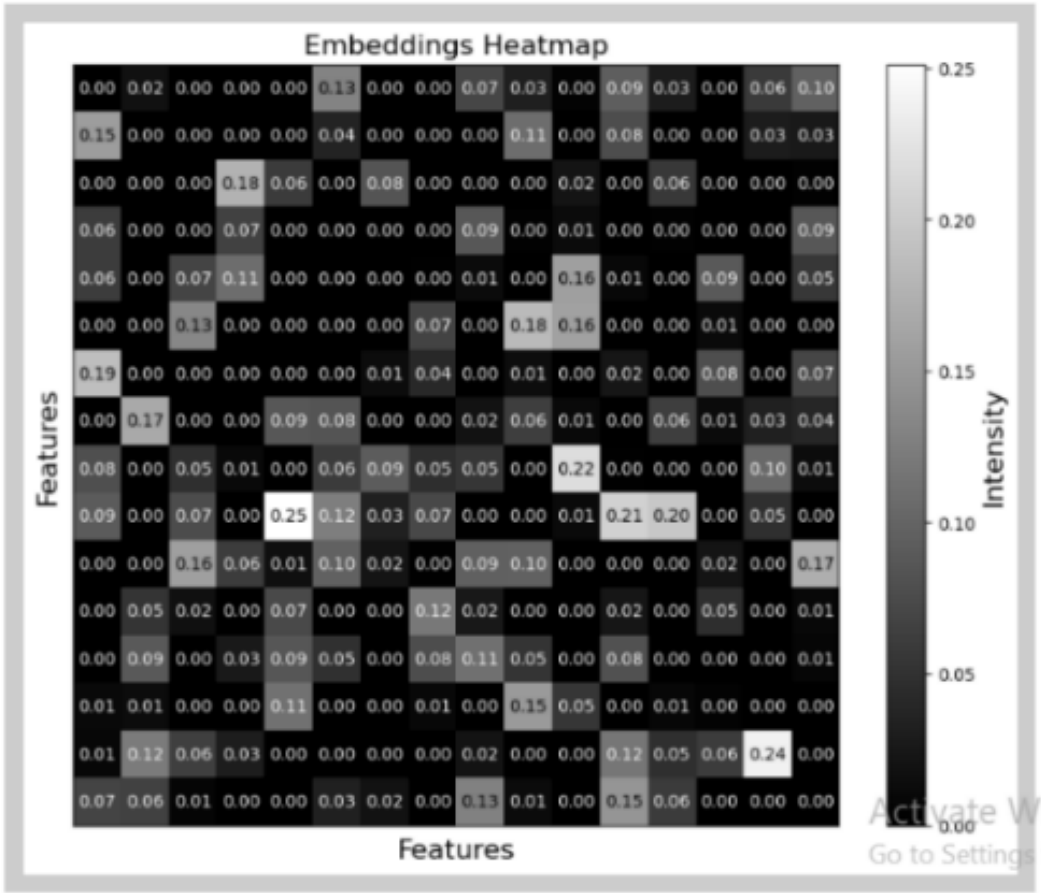
# Results – [9] (Same Text And Transcript)

Text for Male: मेची र महाकाली नदीलाई नेपाली भूभागको प्राकृतिक सिमाना मानिन्छ।

**Original Speaker Embeddings**

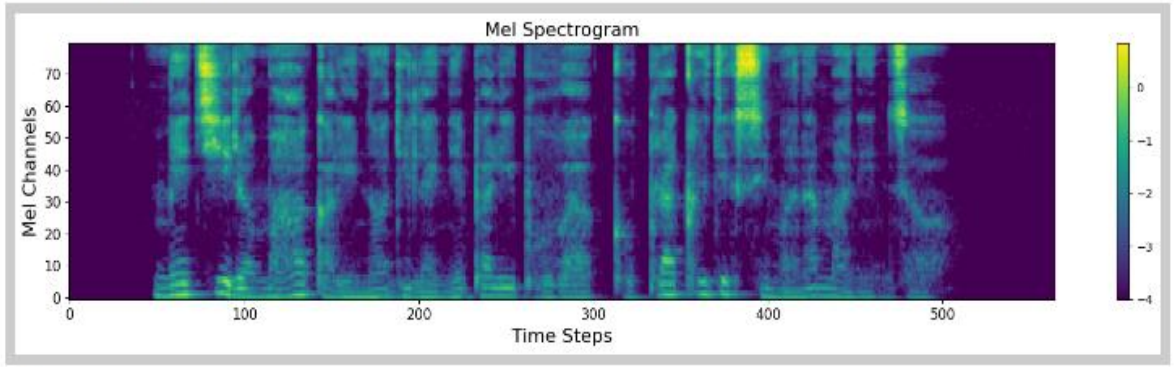



**Cloned Speaker Embeddings**

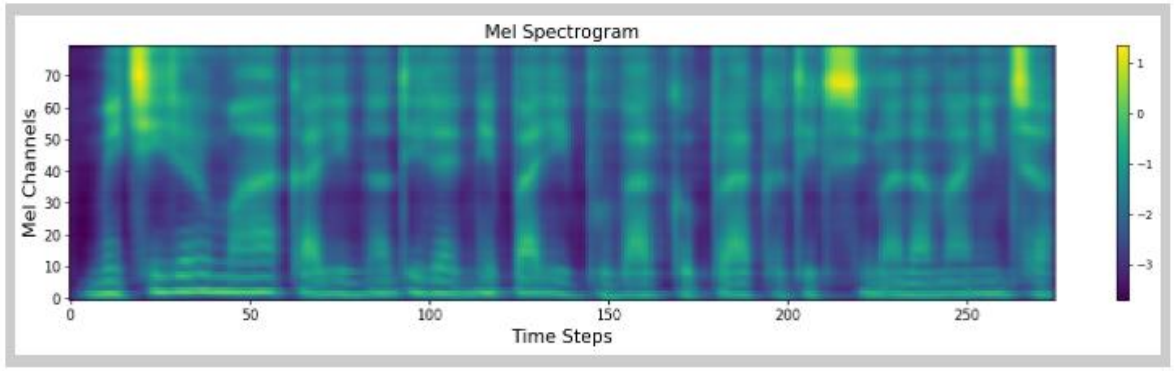


# Results – [10] (Same Text And Transcript)

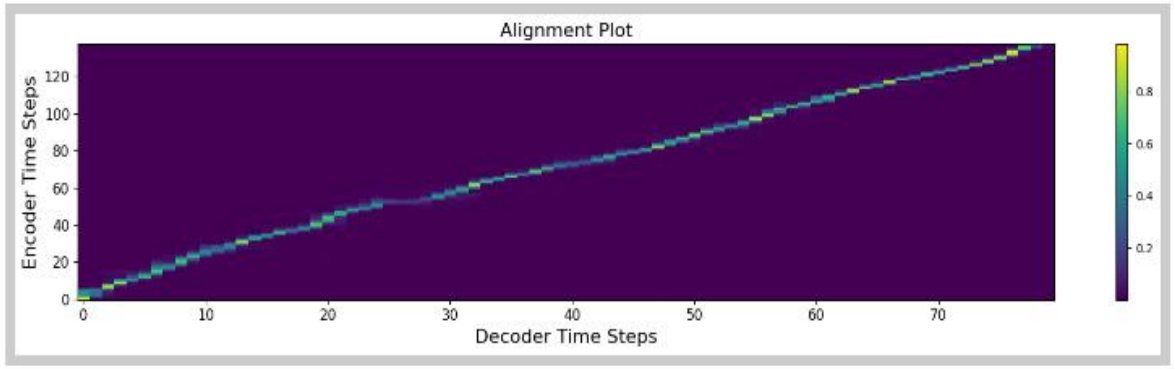
Mel-Spectrogram of Original Audio



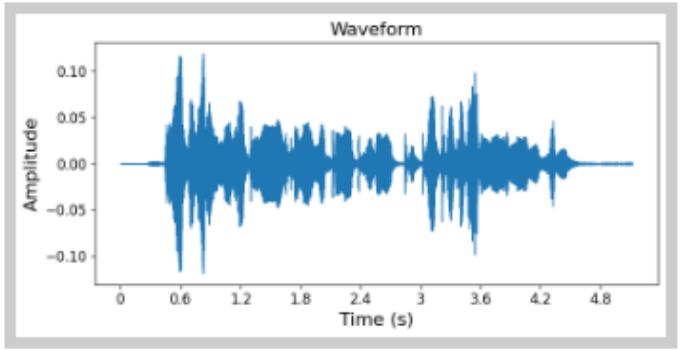

Mel-Spectrogram of Cloned Audio



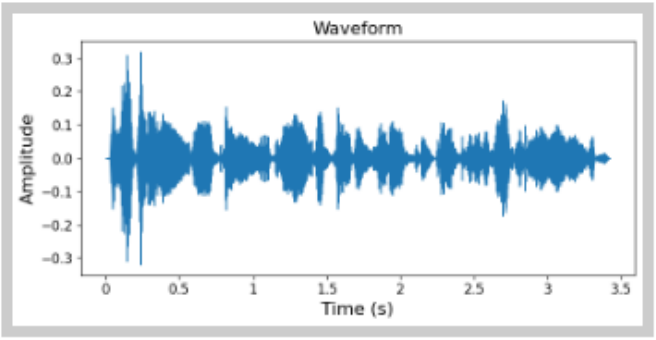

Alignment Plot

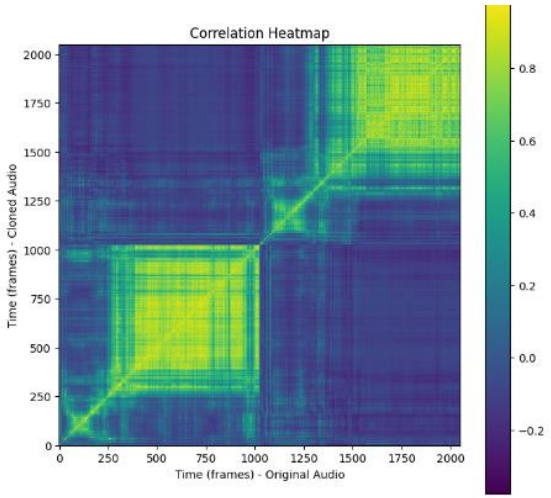



Waveform of Original Audio




Waveform of Cloned Audio



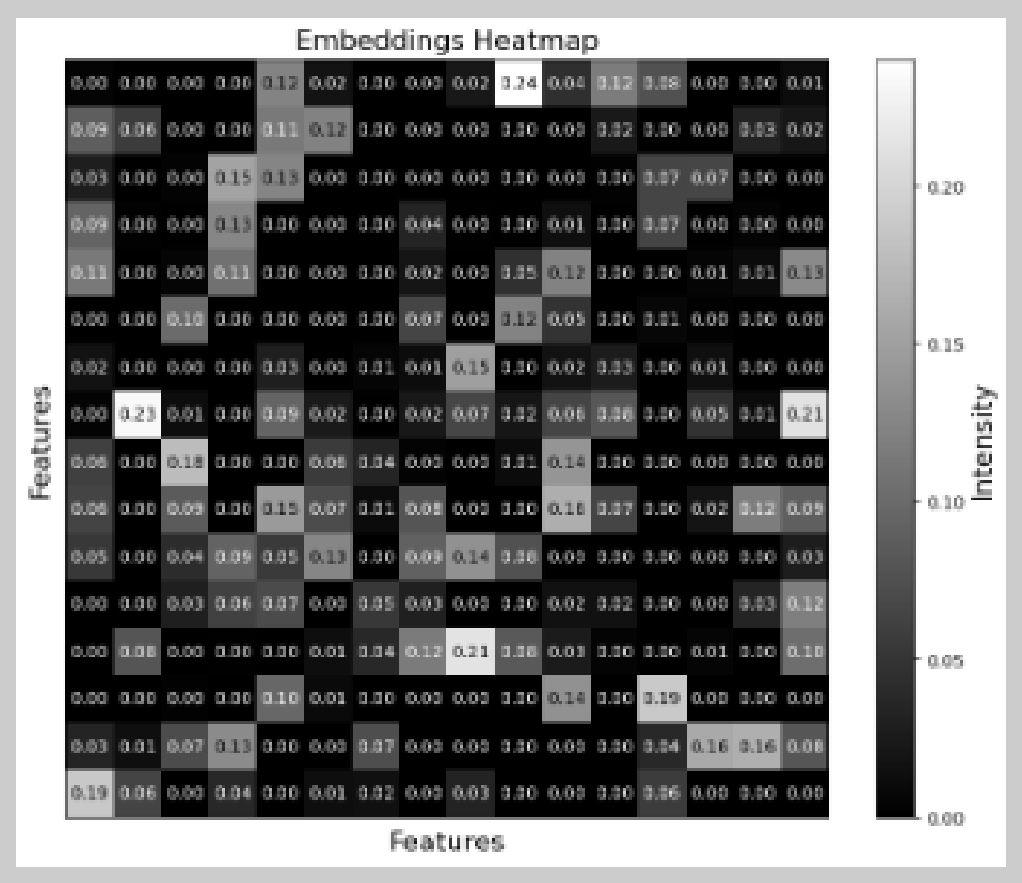



# Results – [11] (Same Text And Transcript)

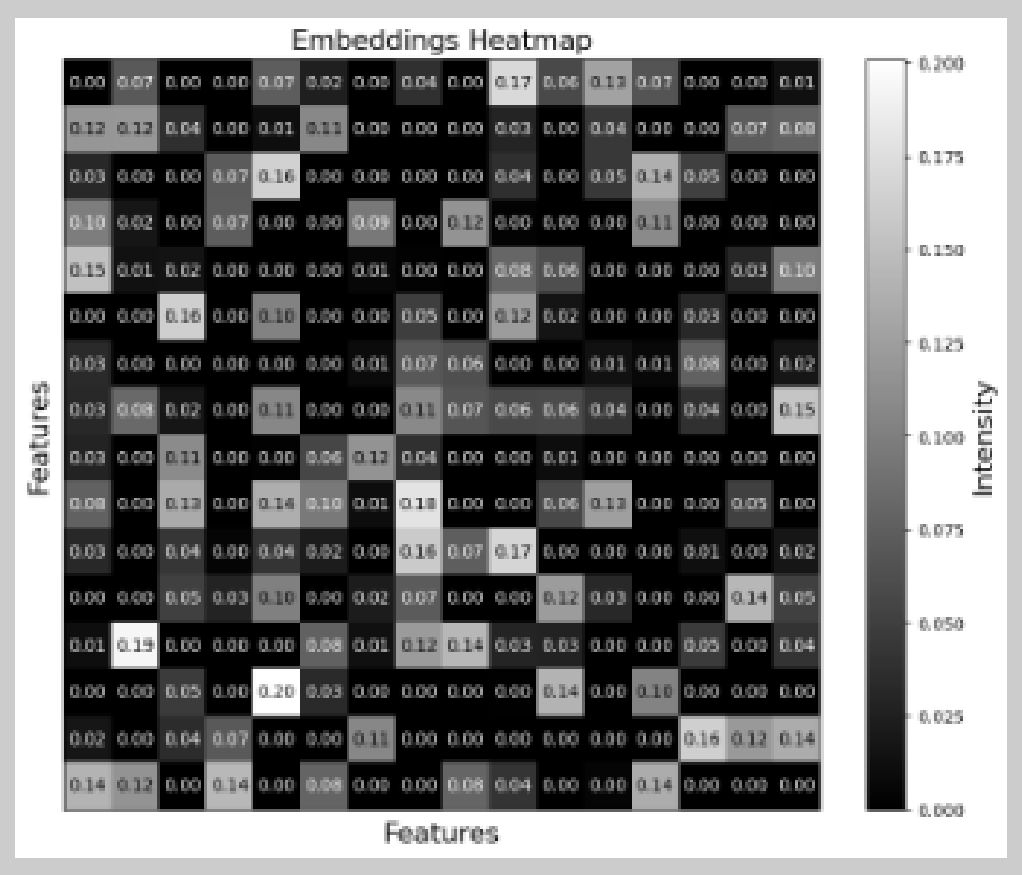
## Text for Female:

उनले कविता र निबन्ध विधामा गरेका योगदानहरू उच्च कोटिको मानिन्छ।

Original Speaker Embeddings

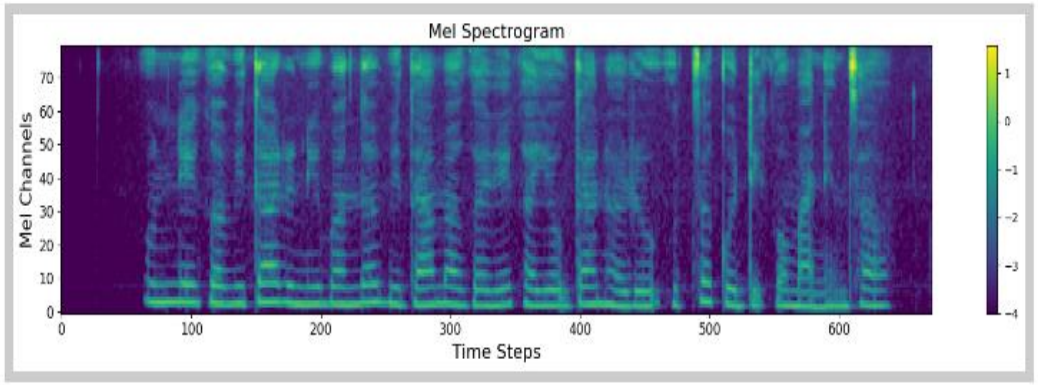



Cloned Speaker Embeddings

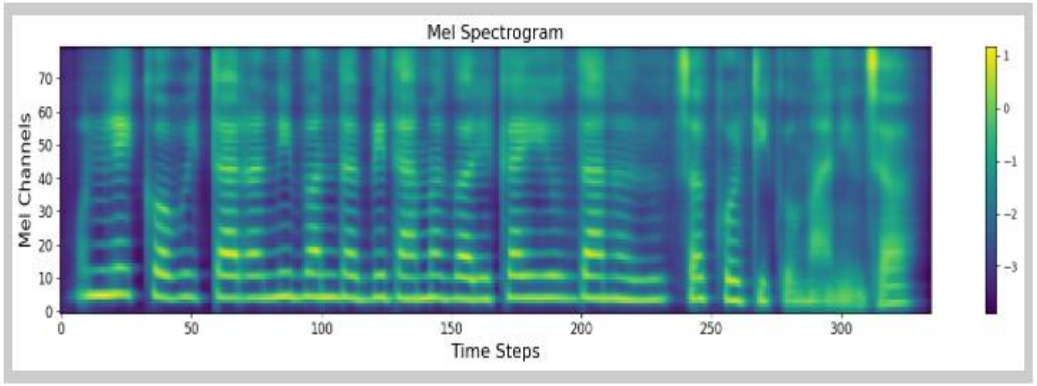


# Results – [12] (Same Text And Transcript)

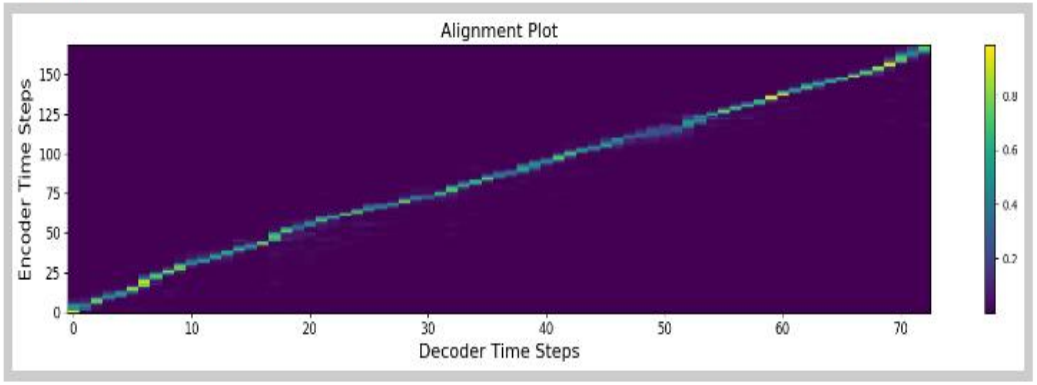
Mel-Spectrogram of Original Audio



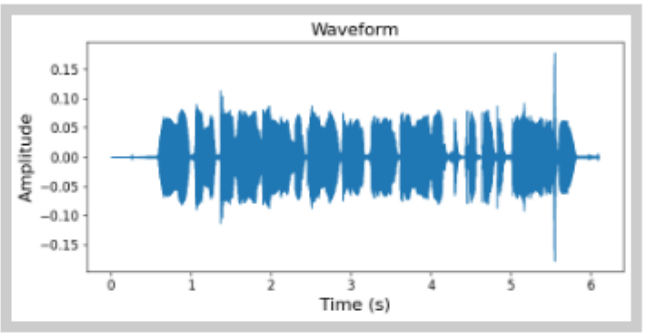

Mel-Spectrogram of Cloned Audio



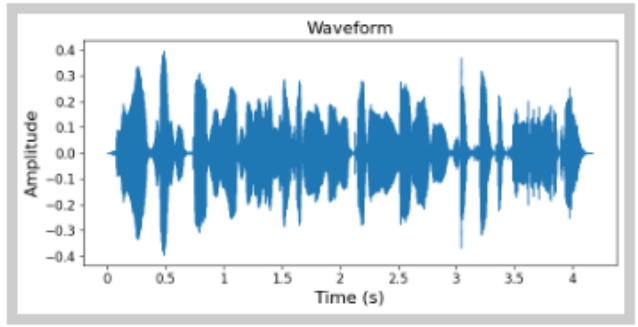

Alignment Plot

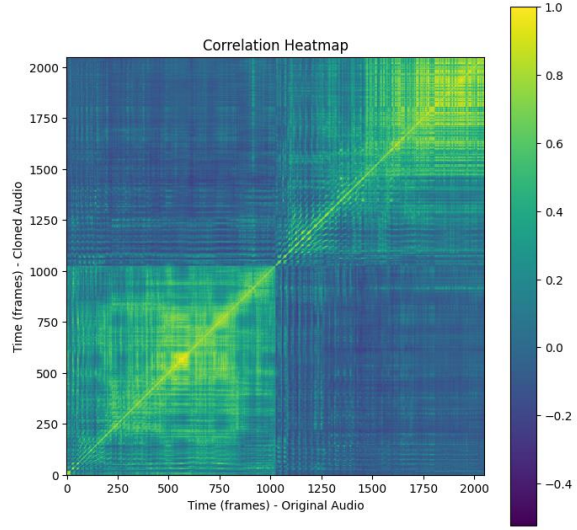



Waveform of Original Audio




Waveform of Cloned Audio



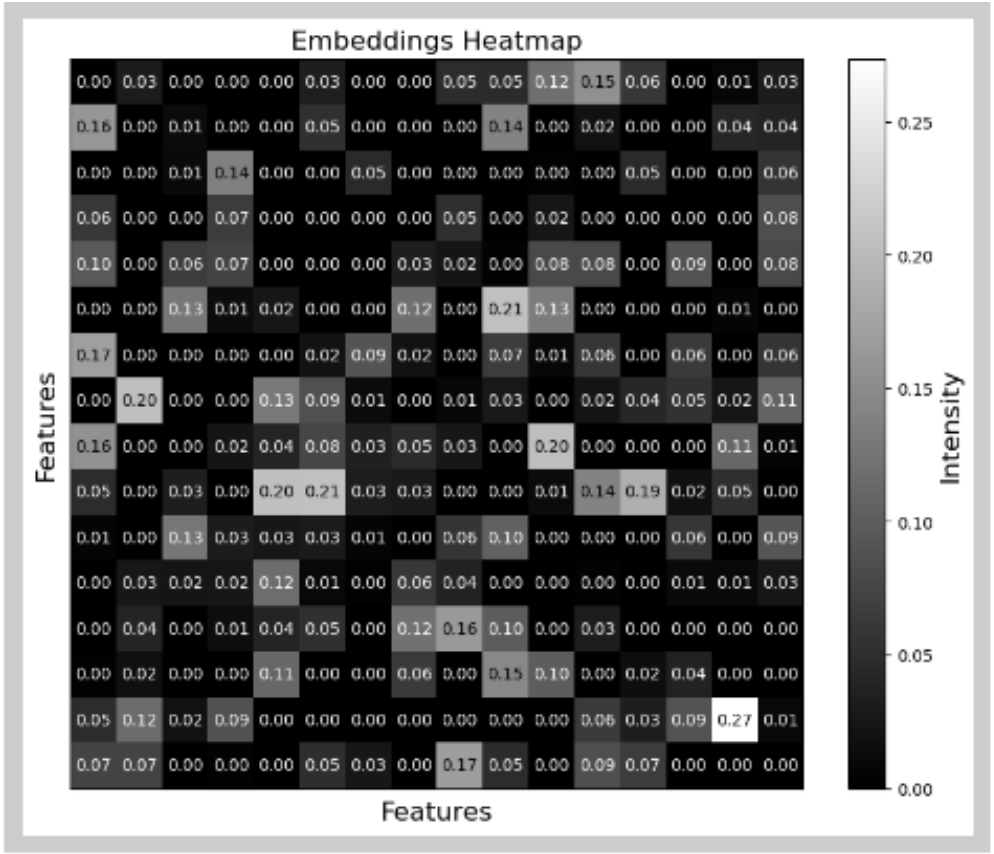



# Results – [13] (Distinct Text & Transcript)

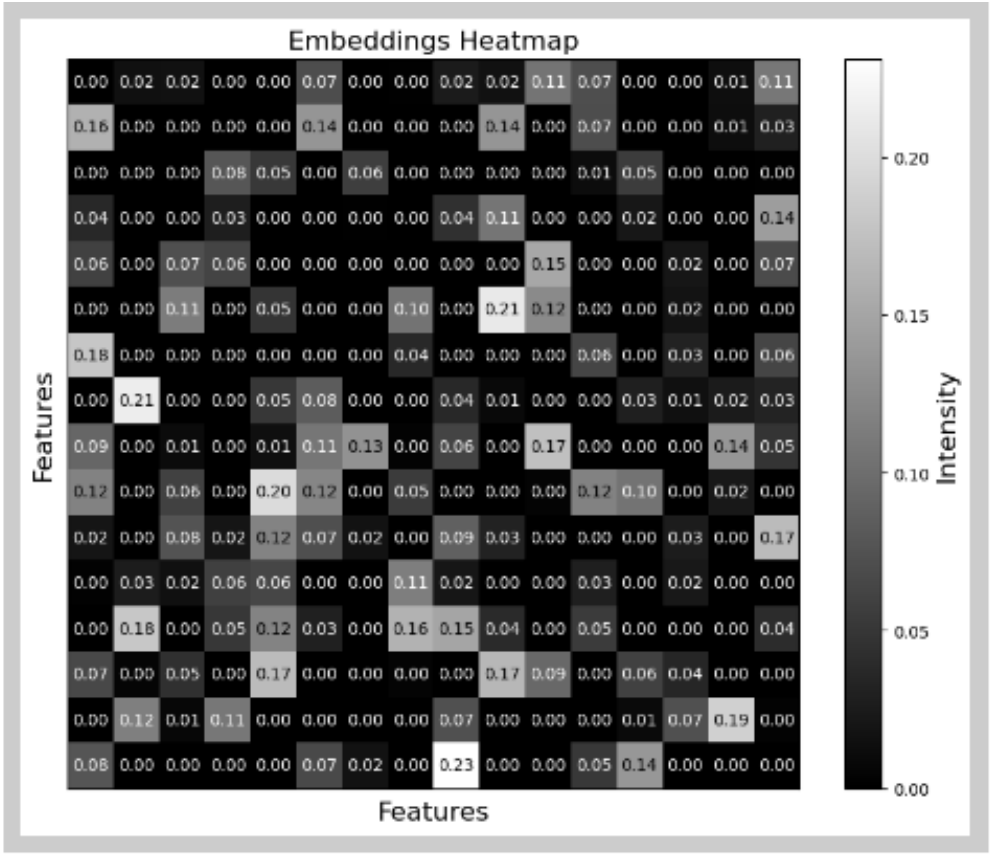
For Male

Input Text: मैले त आलुको सामान्य परिचय मात्र दिन खोजेको हुँ ।

Transcript: मेची र महाकाली नदीलाई नेपाली भूभागको प्राकृतिक सिमाना मानिन्छ।



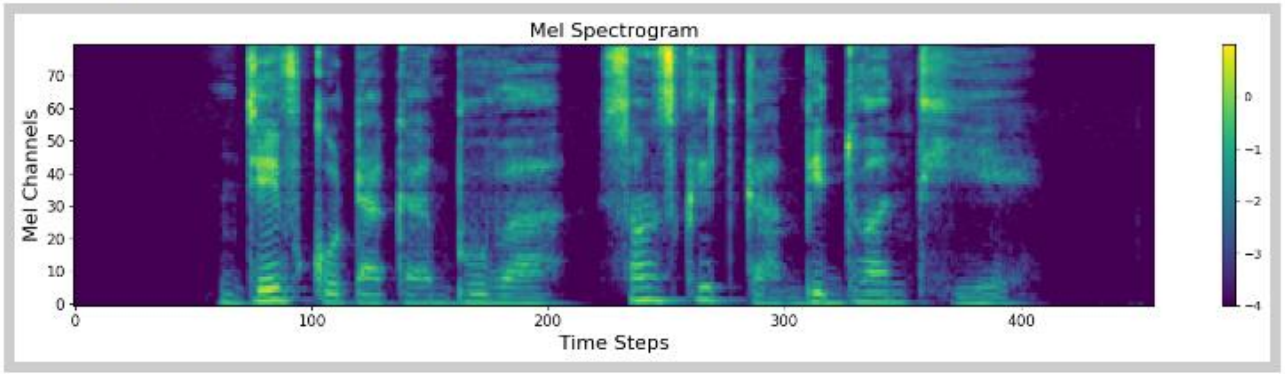

Speaker Embeddings for Transcript



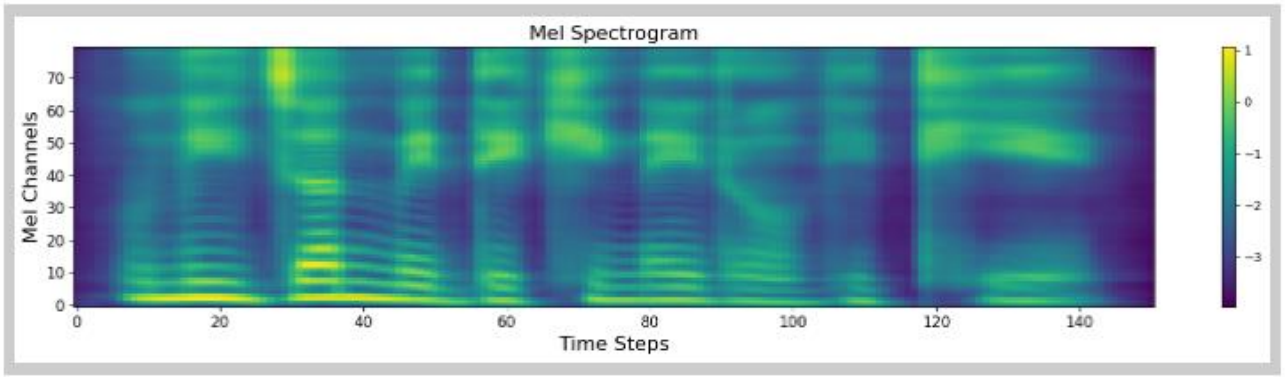

Speaker Embeddings for Input Text

# Results – [14] (Distinct Text & Transcript)

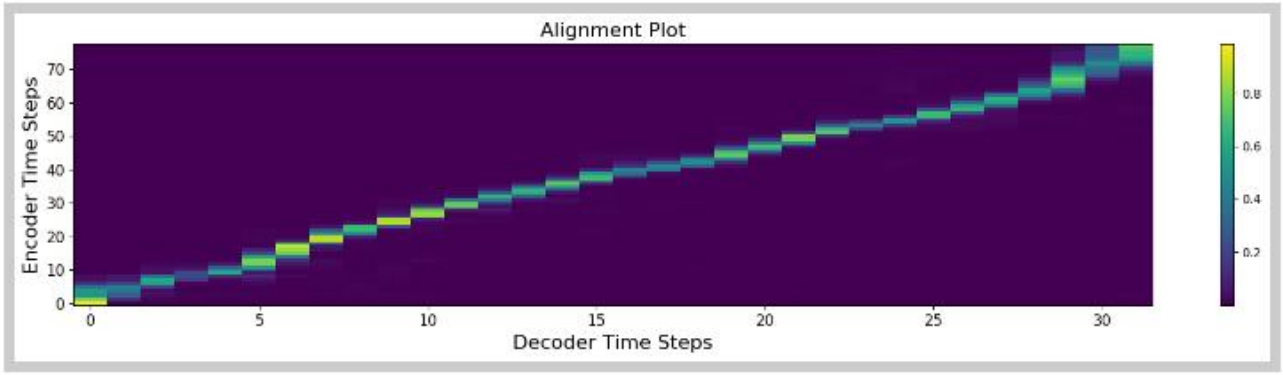
Mel-Spectrogram of Original Audio



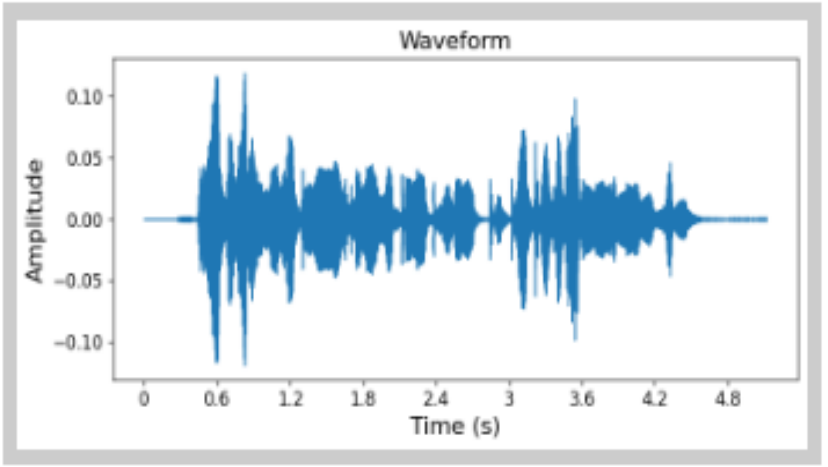

Mel-Spectrogram of Cloned Audio



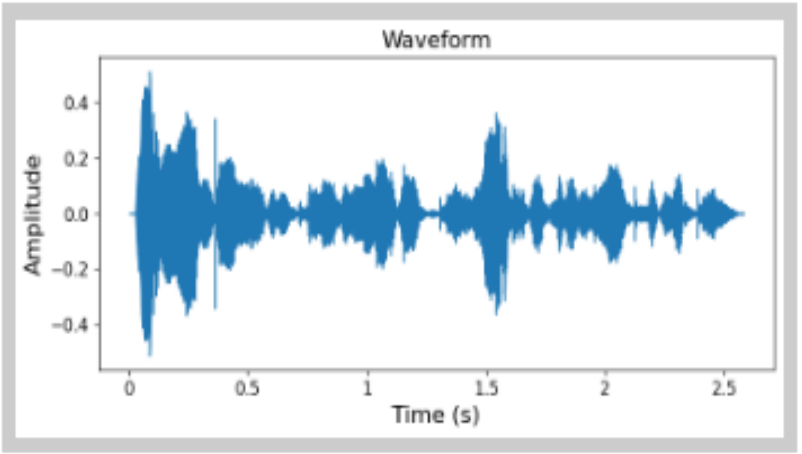

Alignment Plot




Waveform of Original Audio




Waveform of Cloned Audio

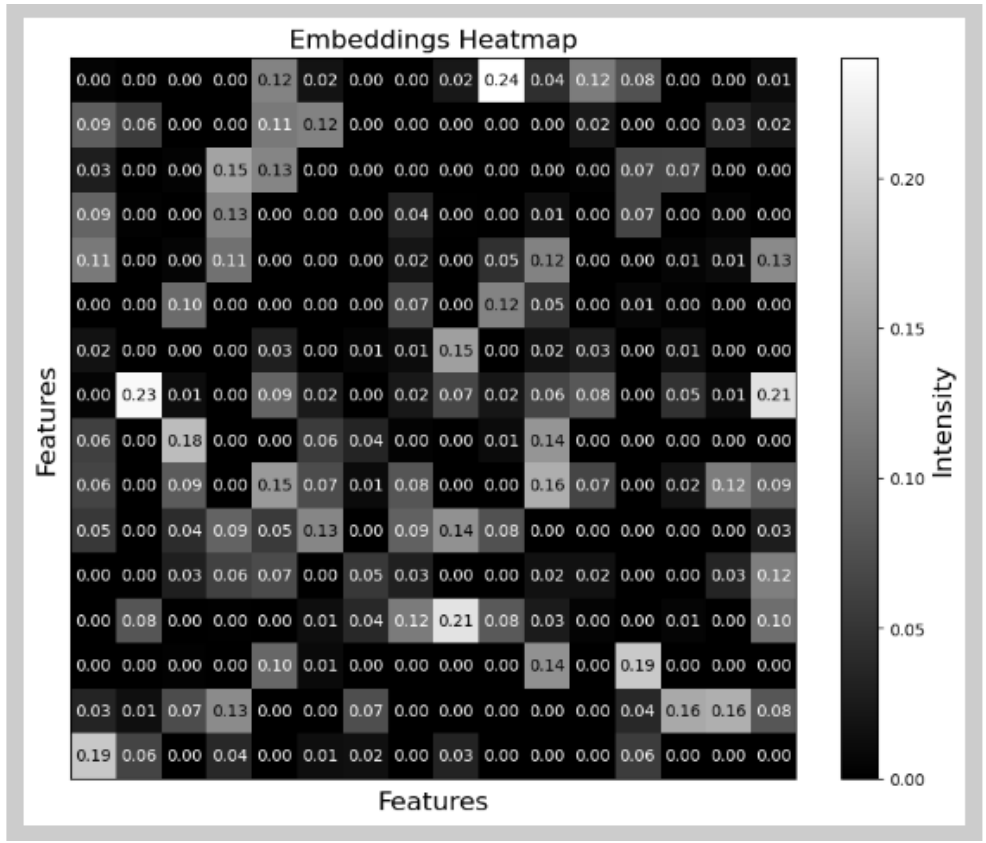


# Results – [15] (Distinct Text & Transcript)

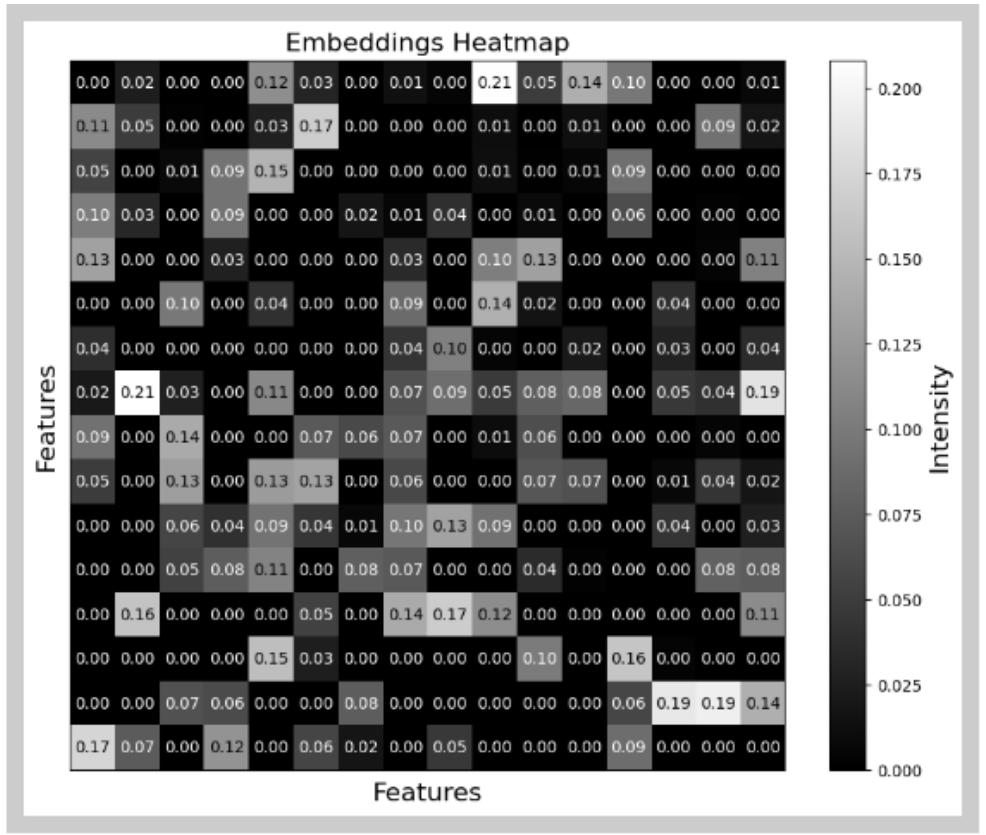
For Female

Input Text: मैले त आलुको सामान्य परिचय मात्र दिन  खोजेको हुँ ।

Transcript: उनले कविता र निबन्ध विधामा गरेका योगदानहरू उच्च कोटिको मानिन्छ।



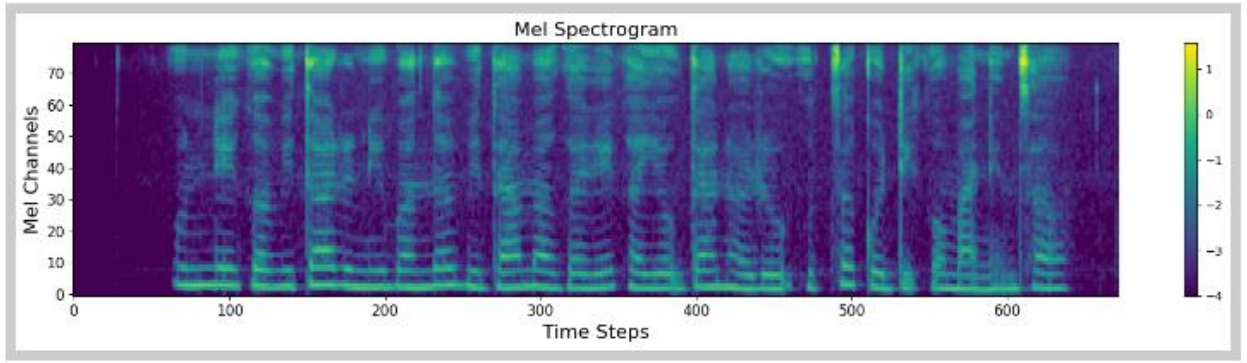

Speaker Embeddings for Transcript



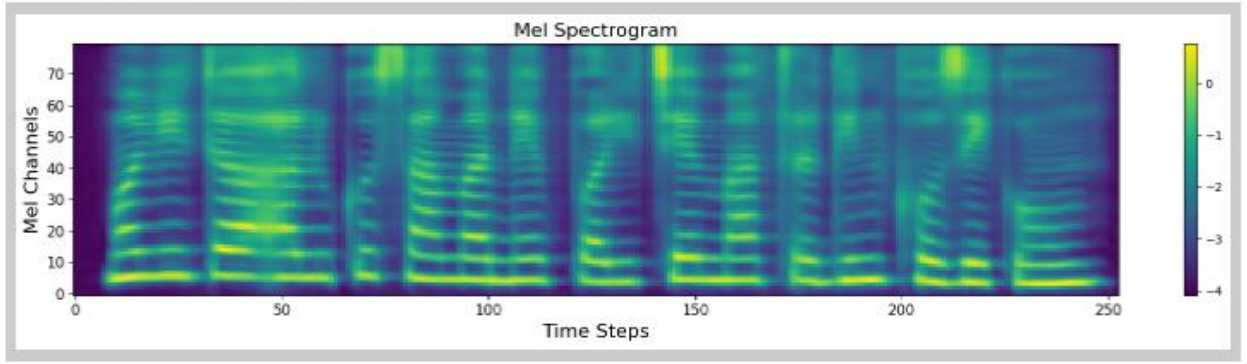

Speaker Embeddings for Input Text

# Results – [16] (Distinct Text & Transcript)

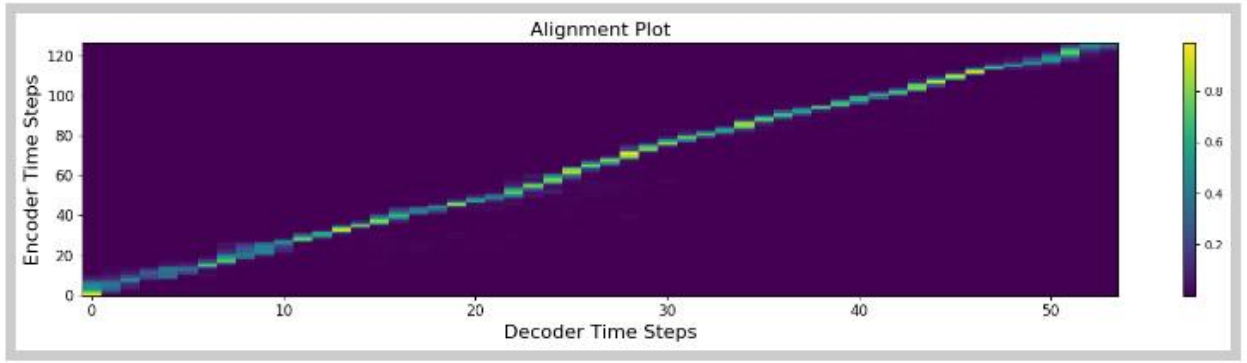
Mel-Spectrogram of Original Audio



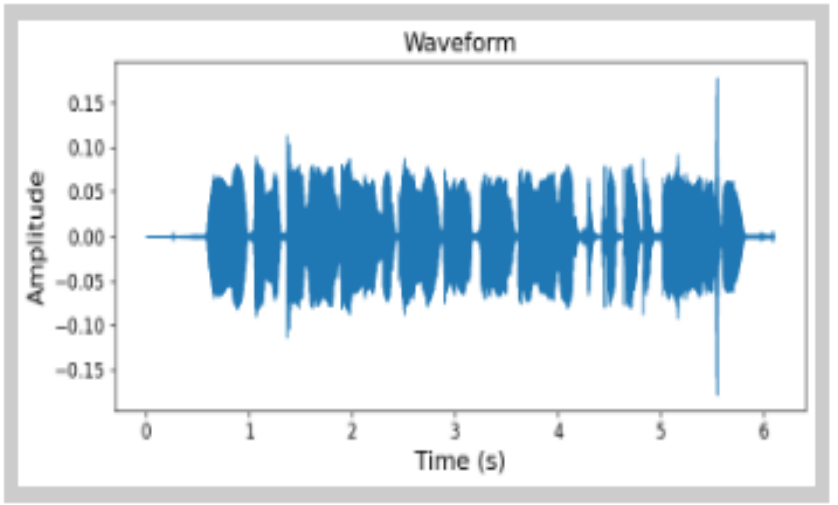

Mel-Spectrogram of Cloned Audio



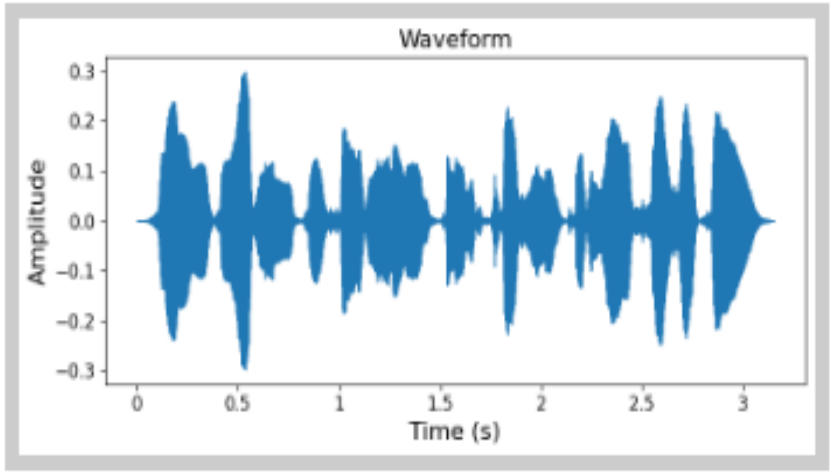

Alignment Plot




Waveform of Original Audio




Waveform of Cloned Audio



# Discussion and Analysis – [1] (MOS Evaluation)

## Voice Cloning MOS Evaluation

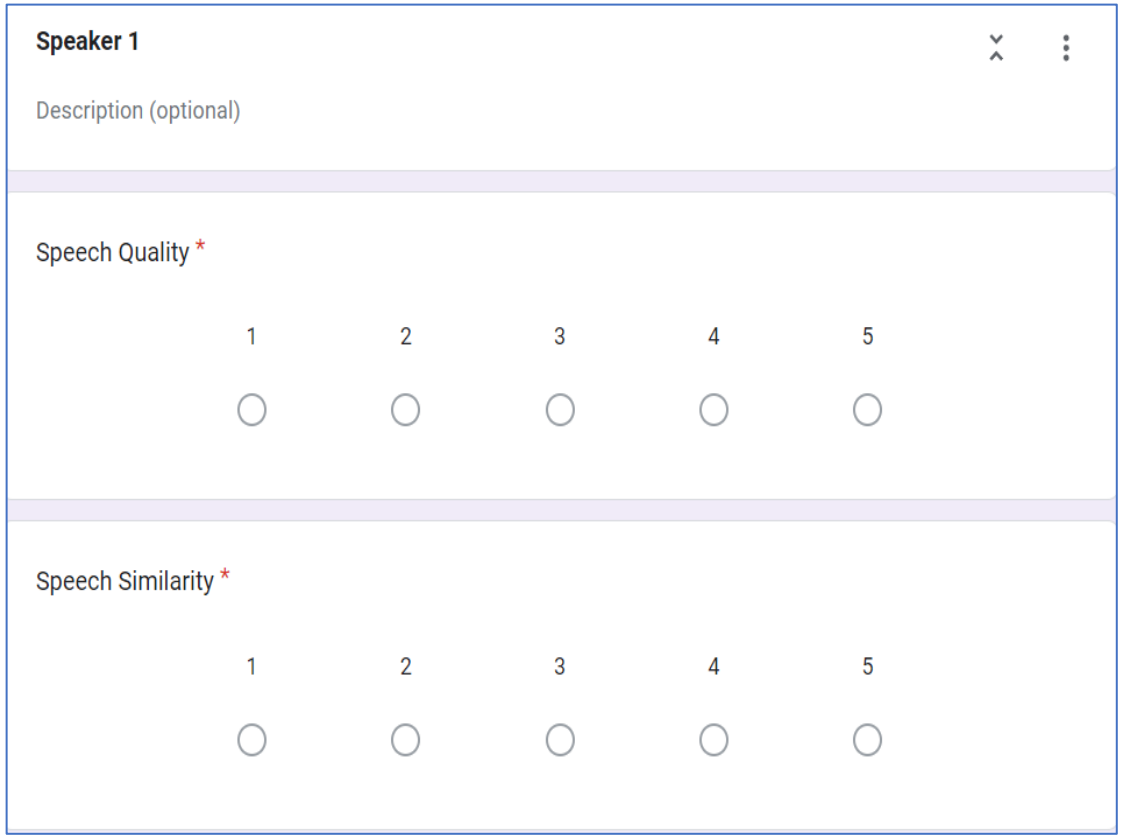
| MOS Score | Speech Quality Description | Speech Similarity Description |
|---|---|---|
| 1 | Not understandable at all | Definitely not the same person, even the gender is different |
| 2 | Some words are unclear and has pronunciation issues | Low chance of being the same person: There is much difference |
| 3 | Generally understandable and acceptable but the rhythmic pause is not good enough. | High chance of being the same person: There is slight similarity. |
| 4 | Natural, clear, and understandable. | Sounds like the same person, but tone and speaking style don't match |
| 5 | Broadcasting level: Unable to distinguish between human voice and synthesized voice | Definitely sounds like the same person: Tone and speaking style match |

## Cloned Audio Samples

NOTE: For each speaker, the left-most audio is the original audio. The other three are the cloned audios.

**Speaker 1**

0:00 / 0:03 0:00 / 0:02 0:00 / 0:03



# Discussion and Analysis – [2] (MOS Evaluation)

| Speaker | Gender | MOS quality | MOS similarity |
|---|---|---|---|
| Speaker1 | Female | 3.18±1.02 | 2.94±1.04 |
| Speaker2 | Female | 3.65±0.82 | 3.33±1.05 |
| Speaker3 | Male | 3.57±0.81 | 3.29±1.18 |
| Speaker4 | Female | 3.26±1,04 | 3.41±1.03 |
| Speaker5 | Female | 3.41±0.95 | 3.43±1.11 |

| Speaker | Gender | MOS quality | MOS similarity |
|---|---|---|---|
| Speaker6 | Female | 3.24±0.95 | 2.91±1.16 |
| Speaker7 | Male | 3.48±0.94 | 3.43±1.11 |
| Speaker8 | Male | 3.15±0.81 | 2.89±1.07 |
| Speaker9 | Female | 3.7±0.90 | 3.50±1.07 |
| Speaker10 | Male | 3.83±0.90 | 3.70±1.16 |

# Discussion and Analysis – [3] (MOS Evaluation)

- Reason for best MOS for speaker 10
  - Basic voice
  - Clear input
  - Pure Nepali words
- Reason for worst MOS for speaker 8
  - Has bass in voice
  - Input being unclear
  - Different linguistic feature
  - Spoken more English than Nepali

# Discussion and Analysis – [4] (Mel-spectrogram Analysis)

Inconsistencies in the original and cloned spectrograms for the case of same text and target audio content.

- Post-Processing Effects
  - Filtering, Dynamic Range Compression, Silence trimming
- Simple model for Speaker Encoder
  - Difficulty capturing all the nuances of linguistic context and prosody
- Artifacts and Distortions in cloned audio

# Discussion and Analysis – [5] (Non-linear Alignment Plot Analysis)

- Variable speaking rate in the target audio
  - During faster speech segments, alignment appear compressed
- Prosody variation
  - Affects timing and duration of speech segments, leading to non-linear alignments
- Phonetic Variability
  - Differences in pronunciation and the transcribed text, especially for language with complex dialects and phonological rules
- Noise and Distortion
  - Background noise, recording artifacts, or signal distortion present in the target audio

# Discussion and Analysis – [6] (Discrepancies in Speaker Embeddings)

- Recording conditions:
  - Example:  Background noise, Microphone issue
- Intraspeaker variability:
  - Example: Emotional and physical state
- Interspeaker variability:
  - Lack of enough training data

# Discussion and Analysis – [7] (Difference in Time Waveform)

Understandable cloned audio despite different time waveform than original because:

- Similar spectral characteristics
  - Example: Energy distribution across different frequency bands
- Preservation of Phonetic Content in Cloned audio
- Robustness of speech perception of humans
  - Preserved key acoustic cues enable speech comprehension.

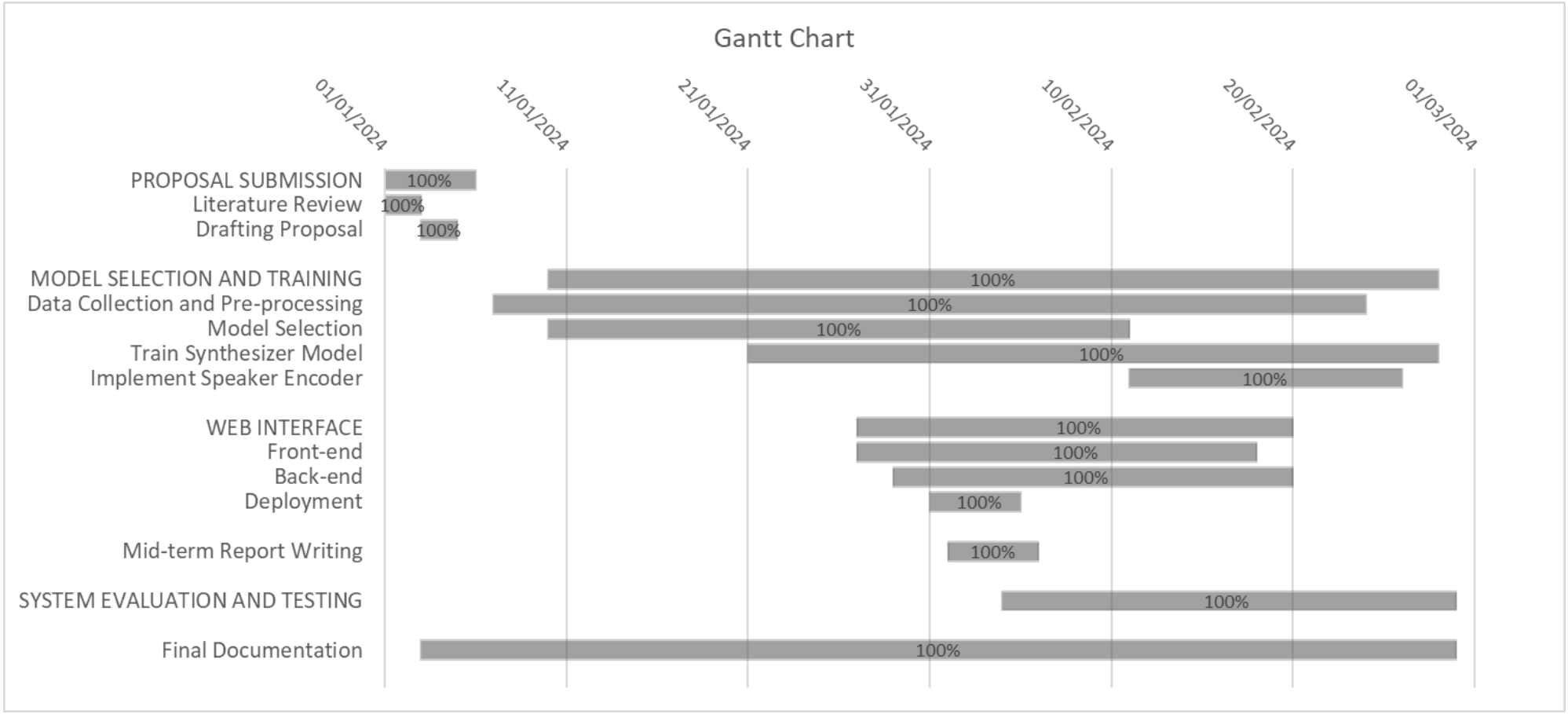
# Future Enhancements

- Additional Speaker for Speaker Encoder
  - Increase number of speaker to ~2000 speakers
- Additional Training Data for Synthesizer
  - Increase the audio and transcript data to ~20 hours
- Usage of Better Vocoder
  - Use latest vocoders like HiFi-GAN or SC-WaveRNN

# Conclusion

- Completion of our objectives
  - Collected and organized datasets for Nepali language
  - Trained multi-speaker generative model for voice cloning
- Stepping-stone for voice cloning in Nepali language
  - Still lacks work to be done compared to English counterpart
  - Middle ground for ML and DSP projects

# Project Timeline



# References – [1]

[1] C.-ping. C. Z.-Sheng. C. Sung-Feng Huang, "PERSONALIZED LIGHTWEIGHT TEXT-TO-SPEECH: VOICE CLONING WITH ADAPTIVE," *IEEE,* no. 05 May 2023, p. 5, 2023.

[2] Y. L. Haitong Zhang, "Improve few-shot voice cloning using multi-modal learning," *IEEE,* no. 18 Mar 2022, p. 5, 2022.

[3] D. Pu, M. Huang, B. Huang, Rui Li, "UNET-TTS: IMPROVING UNSEEN SPEAKER AND STYLE TRANSFER IN," *IEEE,* no. 23 Sep 2021 , p. 6, 2021.

[4] E. Zovato, L. D. Caro, V. Pollet, Giuseppe Ruggiero, "Voice Cloning: a Multi-Speaker Text-to-Speech Synthesis Approach based on Transfer Learning," *arXiv,* no. 10 Feb 2021, pp. 1-5, 2021.

# References – [2]

[5] Y. Chen, L. Chen.  M. Tu, L. Liu, R. Xia, Q. Tian, Y. W. Dongyang Dai, "Cloning one's voice using very limited data in the wild," *IEEE,* no. 7 Oct 2021, pp. 1-5, 2021.

[6] S. Yang, L. Xie, G. Yu, G. Wan, Jian Cong, "Data Efficient Voice Cloning from Noisy Samples with Domain Adversarial Training," *Interspeech 2020,* no. 10 Aug 2020, pp. 1-5, 2020.

[7] J. Kim, J. Bae, Jungil Kong, "HiFi-GAN: Generative Adversarial Networks for Efficient and High Fidelity Speech Synthesis," *ACM,* no. 12 Oct 2020, pp. 1-5, 2020.

[8] J. Chen, K. Peng, W. Ping, Y. Zhou, Sercan O. Arik, "Neural Voice Cloning with a Few Samples," *arXiv,* no. 14 Feb 2018, pp. 1-5, 2018.

# References – [3]

[9] C. Jemine, "Github," 2020. [Online]. Available: https://github.com/CorentinJ/Real-Time-Voice-Cloning. [Accessed 15 February 2024].

[10] A. Taparia, "GeeksForGeeks," GeeksForGeeks, 08 June 2023. [Online]. Available: https://www.geeksforgeeks.org/bidirectional-lstm-in-nlp/. [Accessed 05 Feburary 2024].

[11] L. Wan, Q. Wang, A. Papir, Ignacio Lopez Moreno, "Generalized End-To-End Loss For Speaker Verification" arXiV, p. 2, 2020.

[12] I. Moreno, S. Bengio, N. Shazeer, Georg Heigold, "End-to-End Text-Dependent Speaker Verification," *arXiv,* pp. 3-4, 2015.