Bilingual Speech-to-Speech Translation with Prosody Prediction

Team Members

Pragyan Bhattarai (THA077BEI030)

Prashant Raj Bista (THA077BEI032)

Shakshi Kejriwal (THA077BEI044)

Sudipti Upreti (THA077BEI045)

Supervisor

Er. Kshetraphal Bohara

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

June 2024

Presentation Outline

- Introduction
- Motivation
- Problem statement and Objectives
- Scope
- Limitations
- Application
- Proposed Methodology
- Dataset
- Expected Results
- Timeline
- Budget
- Reference

Introduction

- A translation tool designed to predict prosody along with translation
- A system that translates spoken voice from Nepali to English
- Addresses the challenges of developing a strong and culturally sensitive translation system

Motivation

Problem statement and Objective

Problem Statement:

- Current translation systems lacks to convey the prosody and emotional nuances of spoken Nepali in English

Objective:

- To develop a Nepali-to-English speech-to-speech translation system with prosody prediction on the target language.

Scope

- Designed to develop a Nepali-to-English voice translation with prosody prediction.
- Aims at preserving the emotional aspect, pitch of the speaker, and intensity of speech used by the speaker.

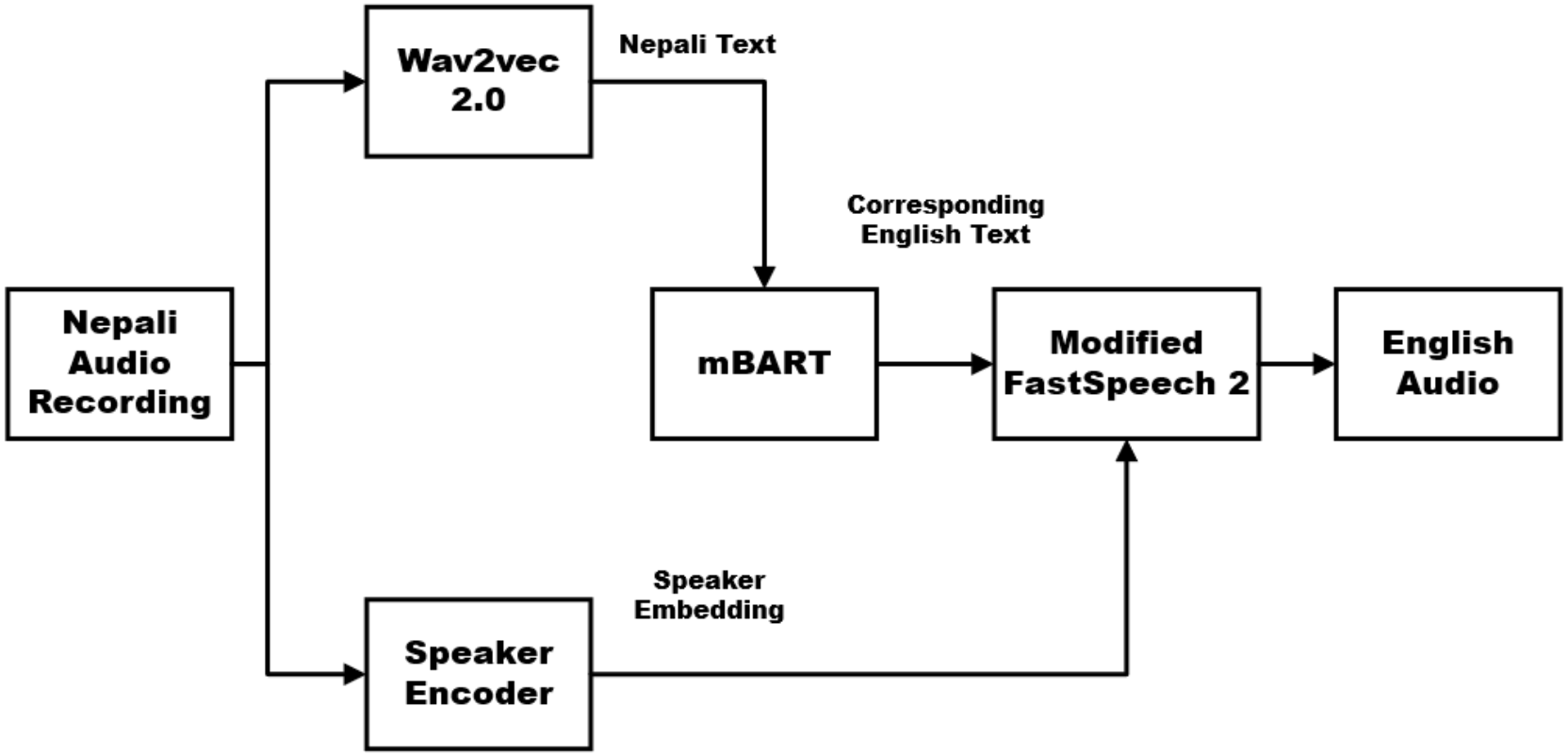
Limitations

- Translates from Nepali to English only
- Not aimed at cross-lingual voice cloning.
- Requires significant computational resources and affect performance on less powerful devices.

Application

- Tourism
- Educational Institutions and Universities
- International Business and Corporate setting
- Government and Public Services

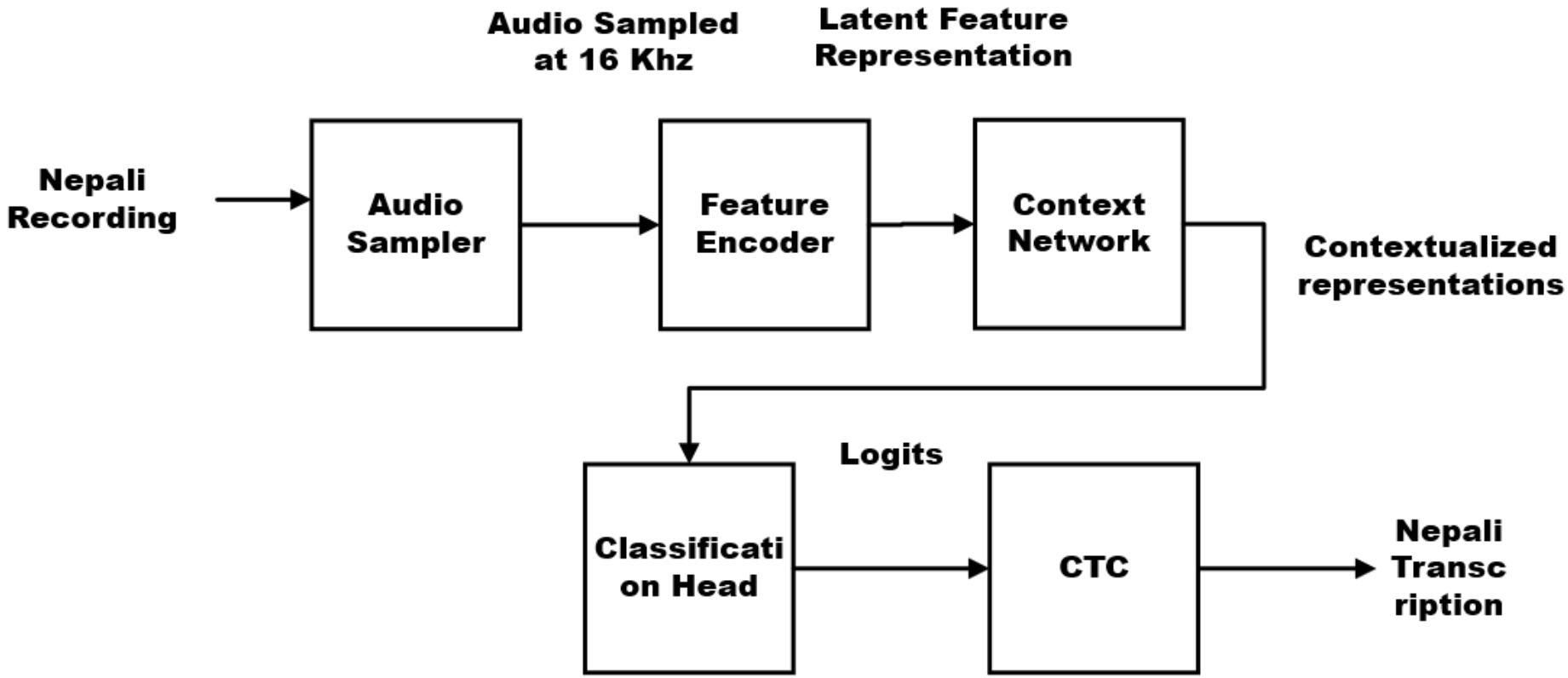
Proposed Methodology - [1]



Proposed Methodology - [2] (Description)

- Recording is first passed to the wav2vec 2.0
- The ASR model provides the corresponding Nepali text as output.
- The output is passed to the mBART that process the Nepali text and present the output with an English Text.
- The English text is passed to the Fast speech 2 along with the speaker embedding from the speaker encoder.
- The modified fast speech produces the corresponding English audio with desired prosody.

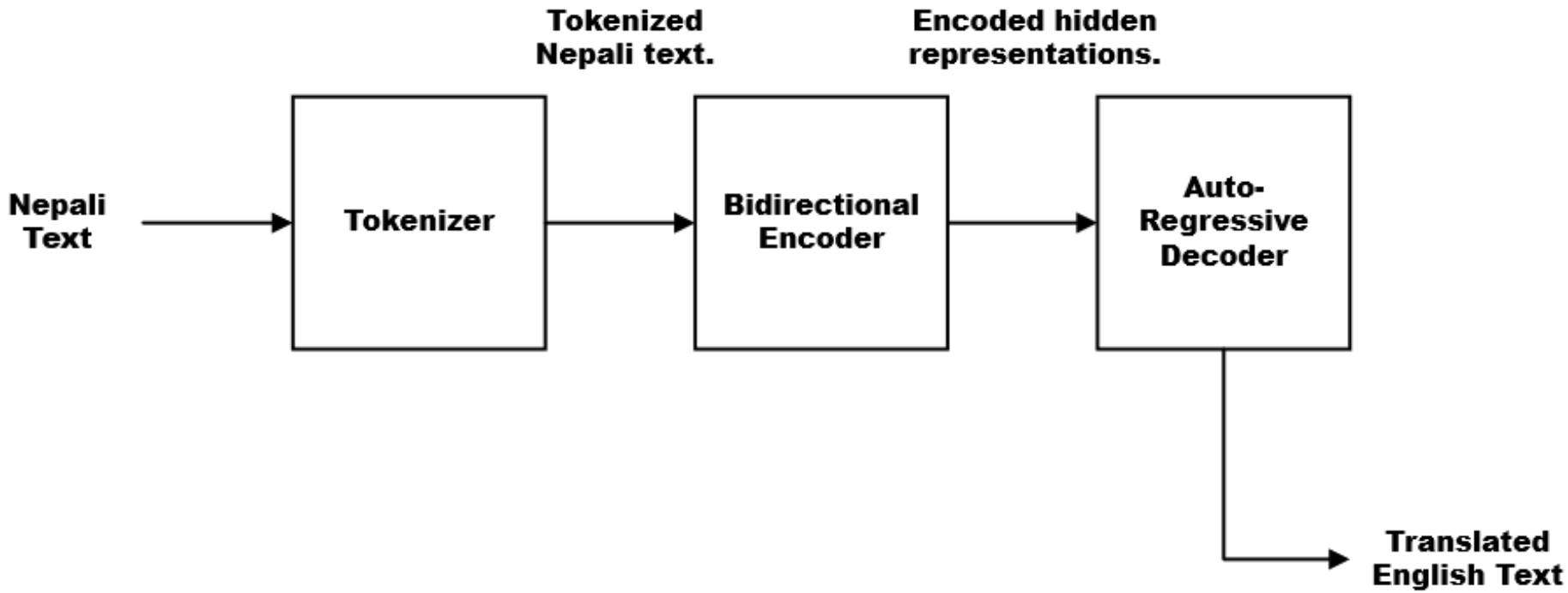
Proposed Methodology - [3] (Data Flow in wav2vec 2.0)



Proposed Methodology - [4] (wav2vec 2.0 Description)

- The recording is first sampled at the rate of 16 kHz.
- The feature encoder extracts the features from the recording and converts the data into the latent feature representation.
- Context network processes and converts the latent feature into contextual representation.
- The classification head converts the contextualized representation into the logits.
- The logits are passed to the CTC which converts the logits into the corresponding Nepali text spoken in the recording

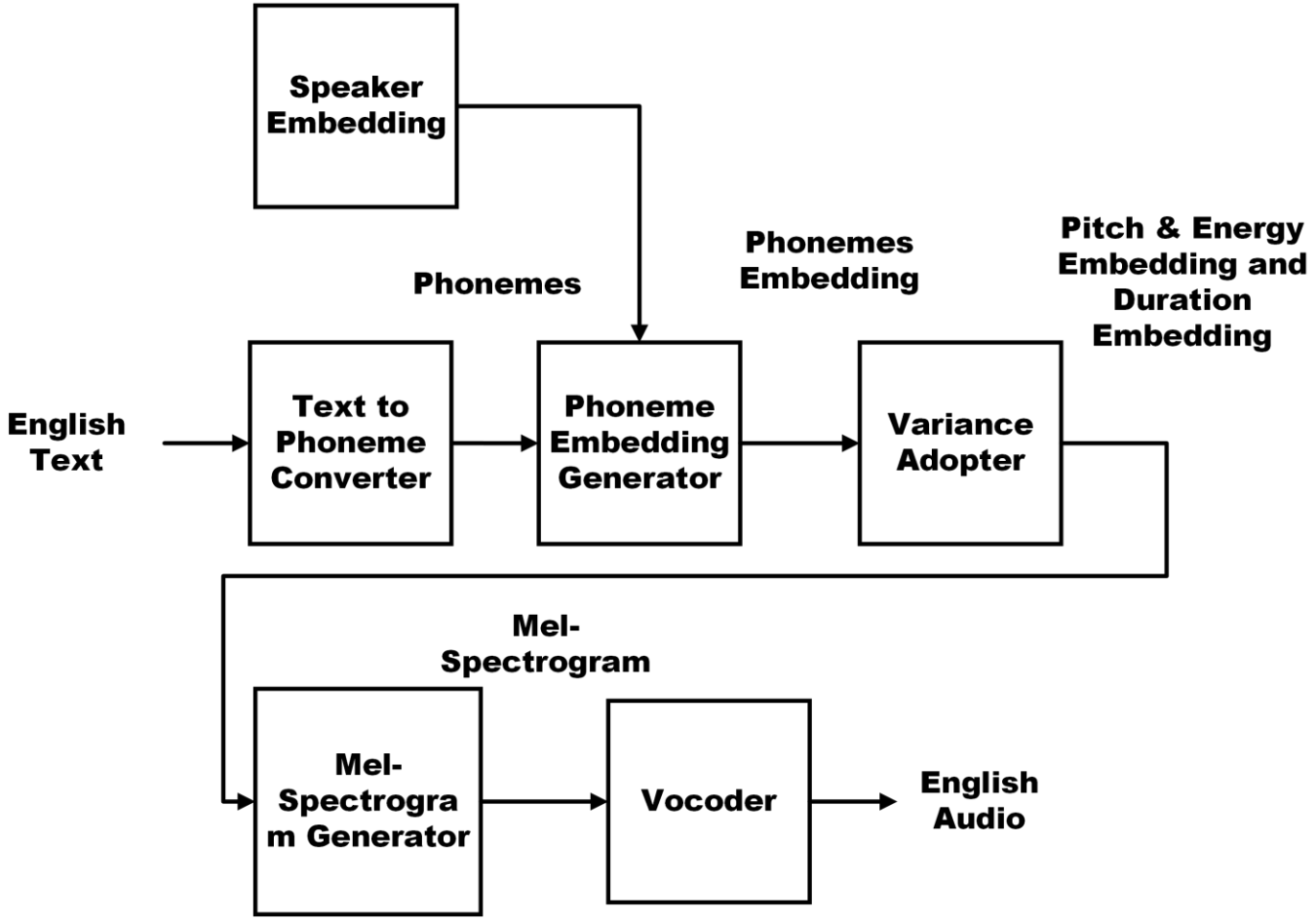
Proposed Methodology - [5] (mBART)



Proposed Methodology - [6] (mBART Description)

- The Nepali Transcription is first passed to the tokenizer that converts the Nepali text into the tokens.
- The tokenized Nepali text is passed to the bidirectional encoder that produces the hidden representation of the data.
- The encoded hidden representation is passed to the autoregressive decoder.
- The decoder produces the output such that the output is transcribed English text.
- The transcribed English text is passed to a TTS model.

Proposed Methodology - [7] (FastSpeech 2)



Proposed Methodology - [8] (FastSpeech 2 Description)

- The English text when passed to the model it is first converted to the respective phonemes.
- The generated phonemes are passed to the phonemes encoder that produces the phoneme embeddings.
- The variance adopter accepts the embedding and predicts the duration, pitch, and energy for the generation of the speech.
- The speaker embedding from the speaker encoder is concatenated with the output of the variance.
- The Mel-Spectrogram generator generates the spectrogram for the speech to be synthesized.

Proposed Methodology - [9] (Speaker Encoder)

- The speaker encoder is a neural network architecture consisting of a LSTM.
- The LSTM captures the dependencies among the speech data.
- The input layer to the encoder is the features that is extracted from the audio such as MFCC, pitch, energy.
- The input features are then normalized, padding, framed into fixed length segments.
- This layer helps in summarizing the information into fixed length vector.
- That will be further be used for concatenation with FastSpeech 2 embedding.

Evaluation Metrices [1] (WER – Word Error Rate)

- WER
  - evaluate the ASR Model.
  - calculates the number of errors divided by total words.

$$WER = \frac{S+I+D}{N}$$

Where S is the number of substituted words, I is the number of intersected words, D is the number of deleted words and N is the total number of words.

Evaluation Metrices [2] (BLEU Score)

- BLEU Score
  - evaluate the Machine Translation Model.
  - measures similarity between the predicted translation and reference translation by computing the overlap between output and reference text.

$$\text{Brevity Penalty } p = \begin{cases} 1 \; if \; c > r \\ e^{1-\frac{r}{c}} \; if \; c \leq r \end{cases}$$

$$BLEU = p \times e^{\sum_{n=1}^{N}(\frac{1}{N} * \log P_n)}$$

Evaluation Metrices [3]
(MOS – Mean Opinion Score)

- MOS
  - evaluate the quality of audio synthesized.
  - defined in the range of 1 to 5 as,

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

Where R_n is the rating by each person and N is the total number of people participated

Dataset - [1]
(Available Dataset)

Dataset for ASR	Dataset for NMT	Dataset for TTS
Common Voice by Mozilla	Opus Corpora	LJ Speech Dataset
Open SLR o SLR 143 o SLR 43 o SLR 54		
Nepali STT dataset on Kaggle		

Dataset - [2]
(Dataset for ASR)

S.N.	Identifier	Duration (in hrs)	No. of Speakers	Audio Samples
1	SLR43	2.47	2	666
2	SLR143	1.2	19	2064
3	SLR54	165	527	157905
4	Parliament Speech collected by Ishwor Subedi	2.82	-	339
5	Common Voice By Mozilla	18	169	7110

Dataset - [3]
(Dataset for MT and TTS)

Dataset for MT

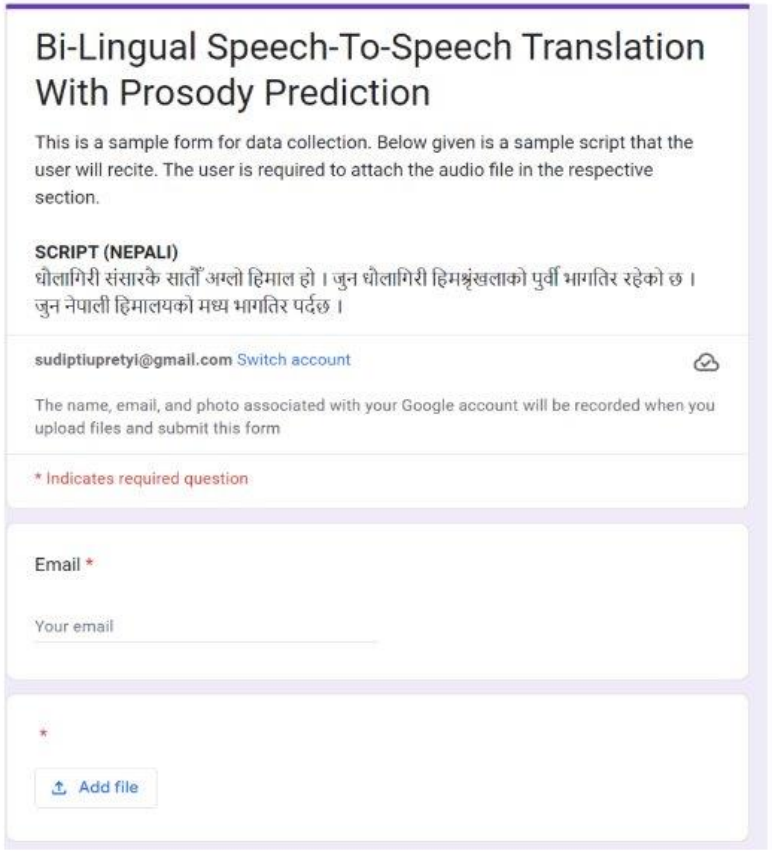
S.N.	Identifier	Ne-En Sentence Pairs	Nepali Tokens	English Tokens
1	Opus Corpora	21,807,645	196,405,524	196,345,204

Dataset for TTS

S.N.	Identifier	Duration (in hrs)	No. of Speakers	Audio Samples	Transcripted Words
1	LJ Speech	24	1	13,100	225,715

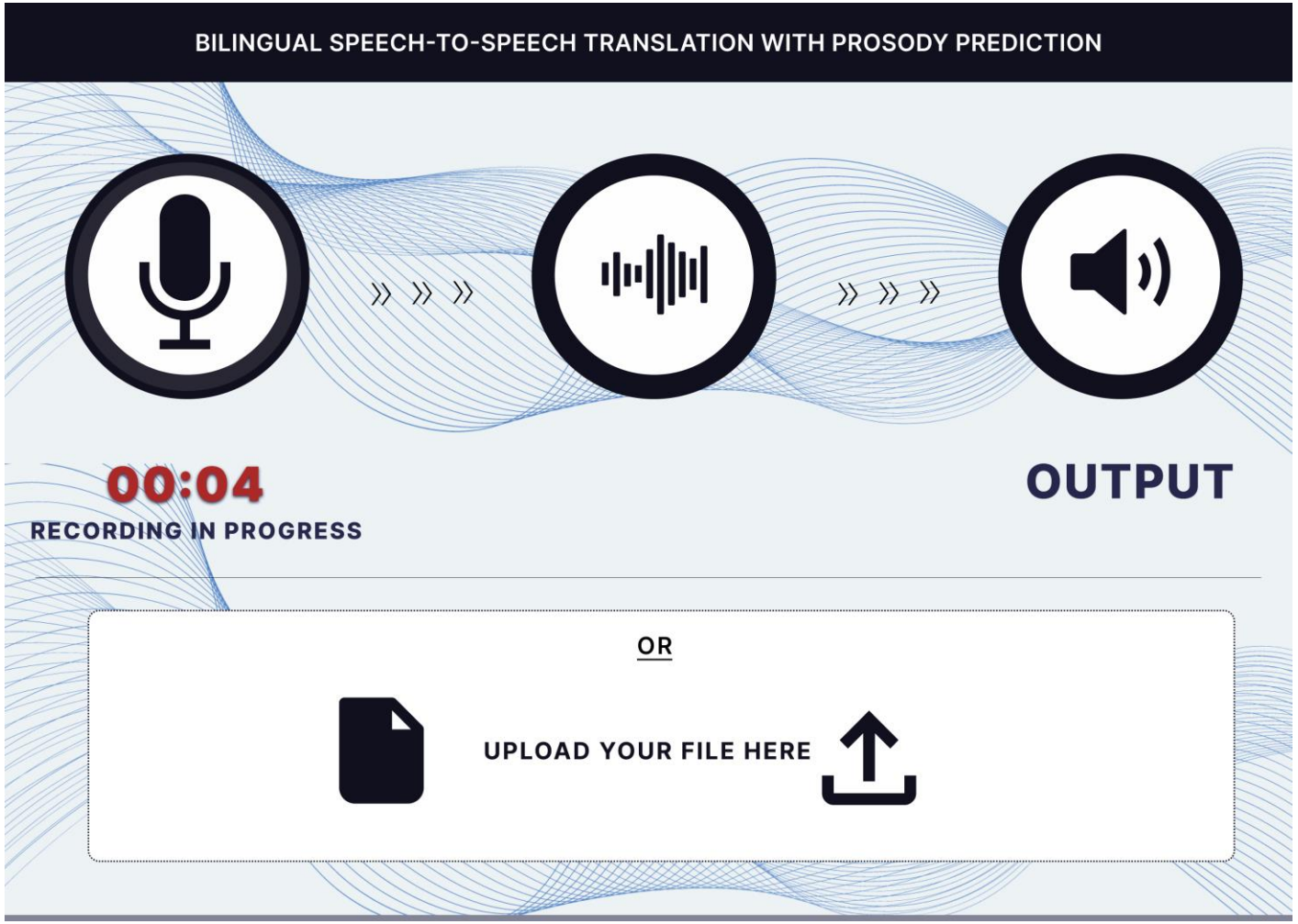
Dataset - [4]
(Proposed Method for Dataset Collection)

- Google Forms



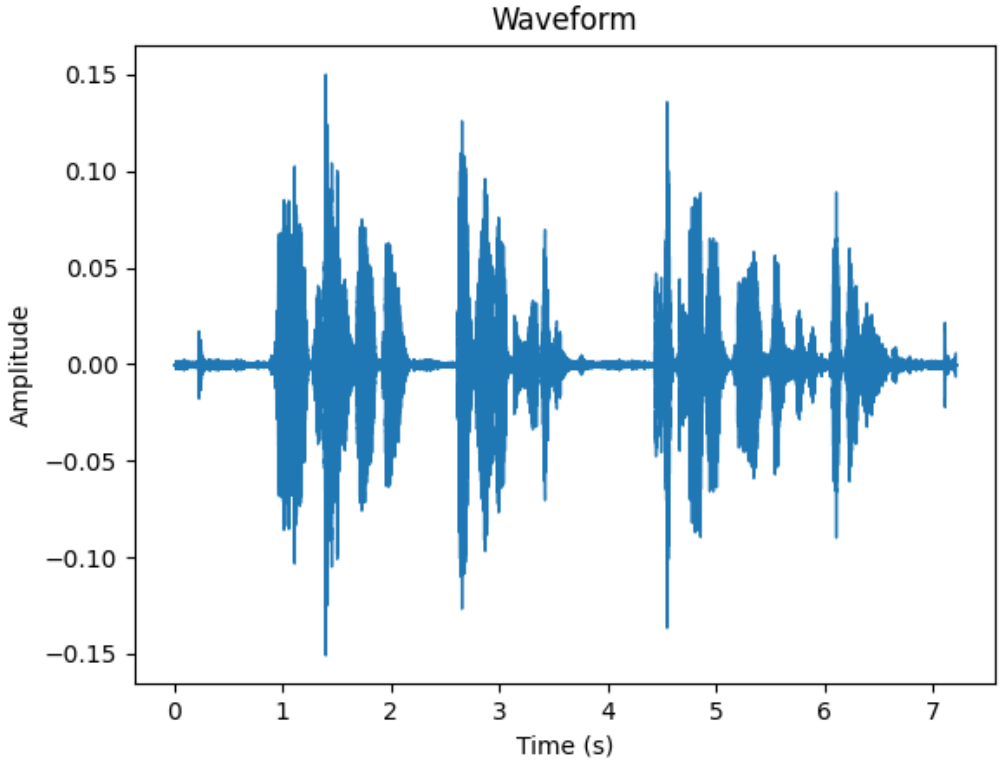
- Website for Data Collection

Expected Results - [1]



Expected Results - [2]

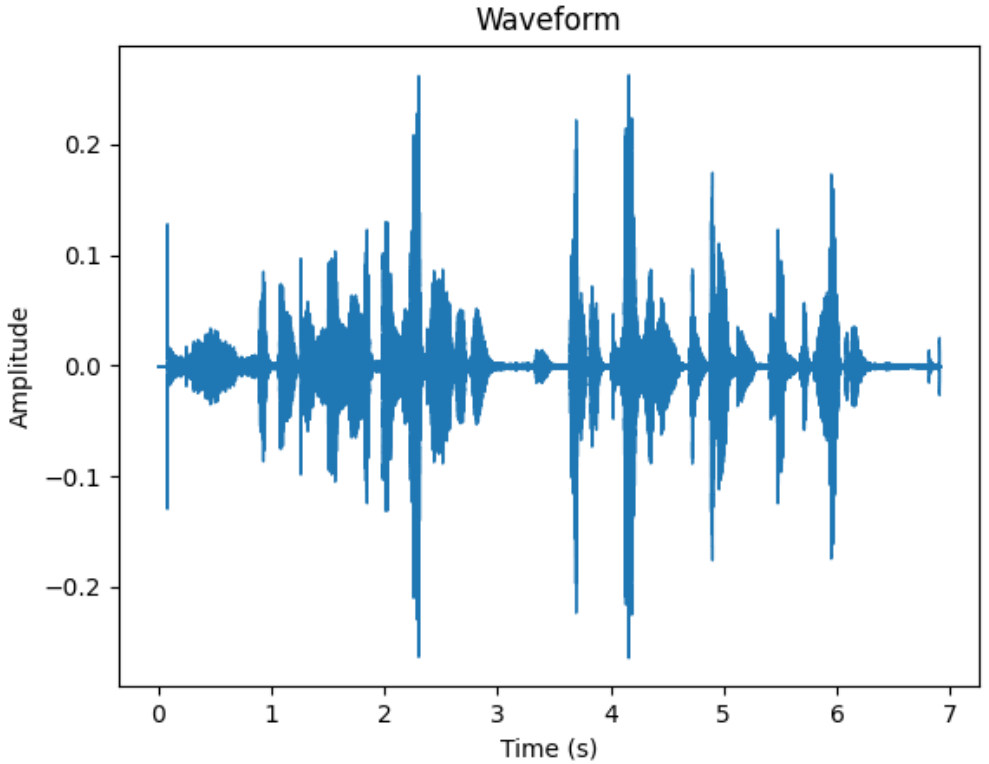
Nepali Script : “यहाँको सुन्दरता अवर्णनीय छ। कस्तो शान्त र मनमोहक वातावरण।”



Ground Truth Nepali Waveform

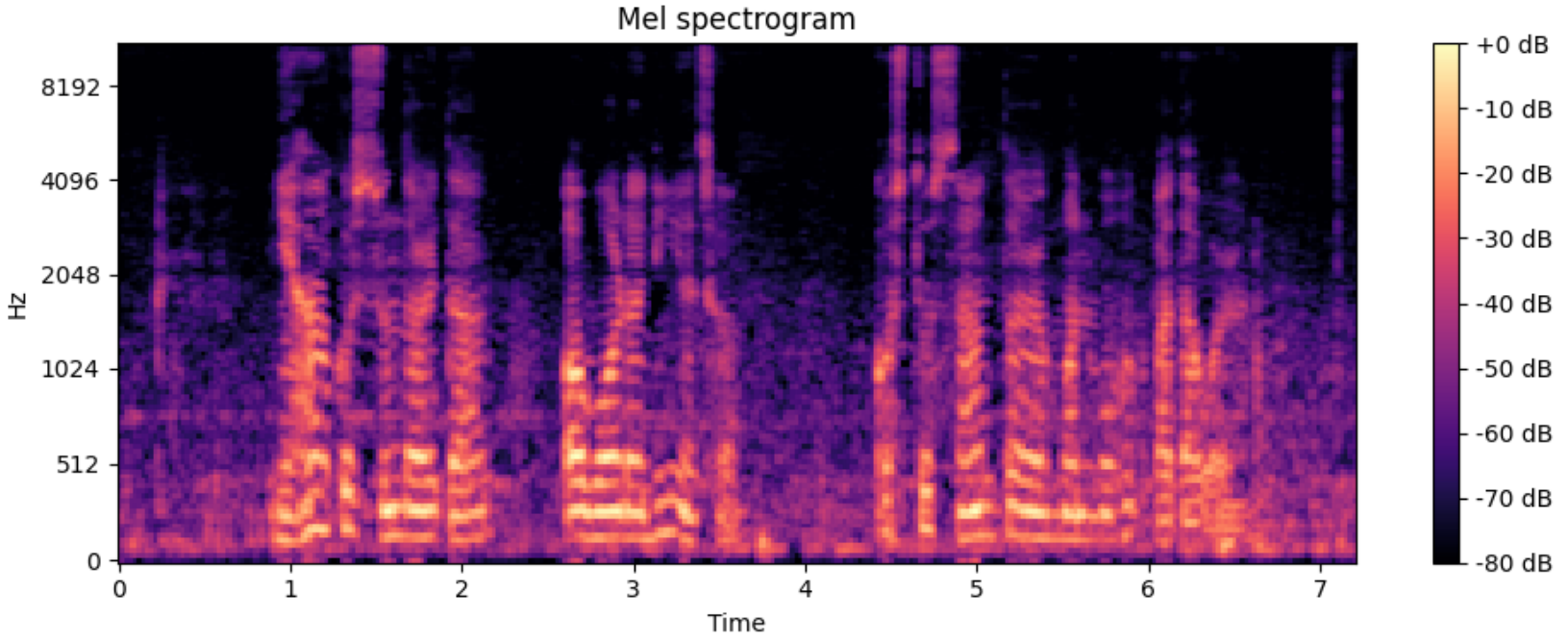
Expected Results - [3]

English Script : “The beauty here is indescribable. Such a peaceful and enchanting atmosphere. ”



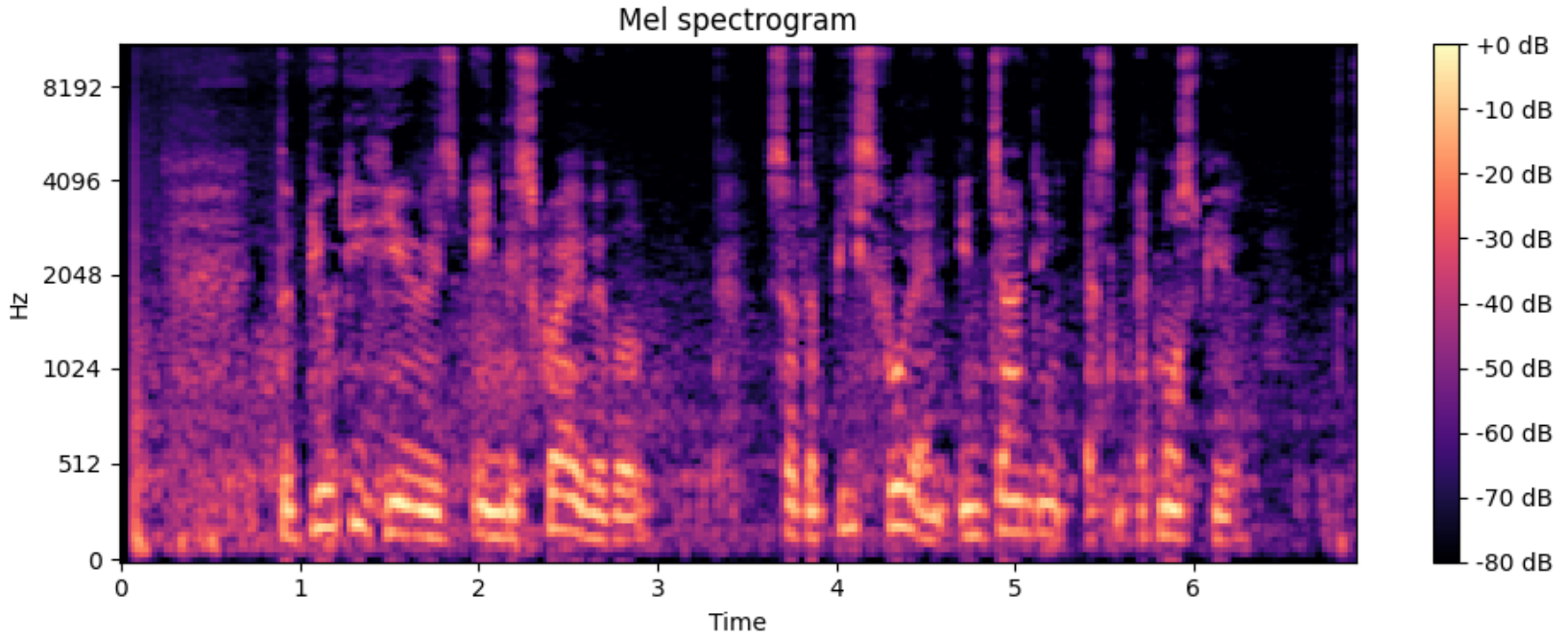
Predicted English Waveform

Expected Results - [3]



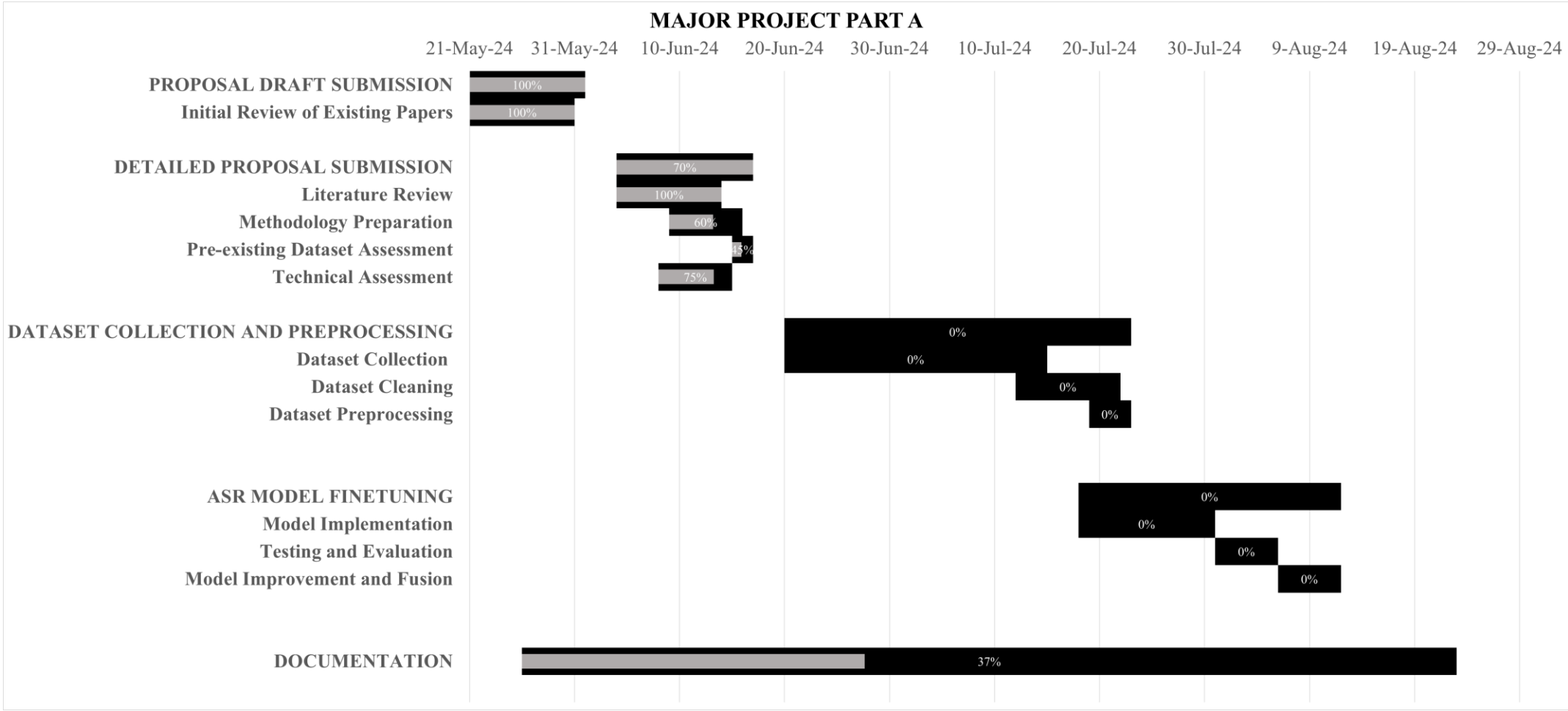
Mel-Spectrogram of Nepali Speech

Expected Results - [4]

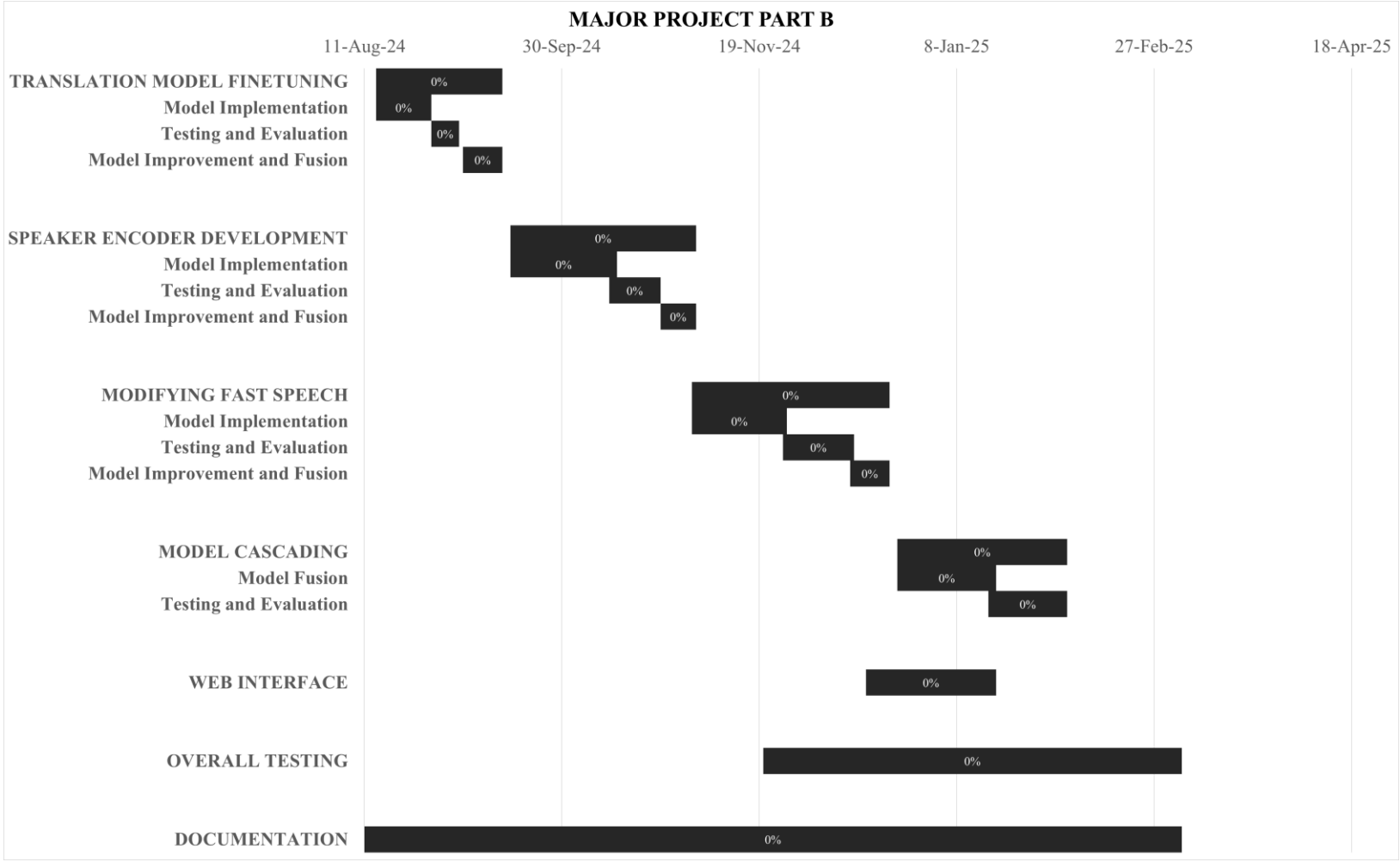


Mel-Spectrogram of English Speech

Timeline - [1]


MAJOR PROJECT PART A
21-May-24
31-May-24
10-Jun-24
20-Jun-24
30-Jun-24
10-Jul-24
20-Jul-24
30-Jul-24
9-Aug-24
19-Aug-24
29-Aug-24
PROPOSAL DRAFT SUBMISSION
100%
Initial Review of Existing Papers
100%
DETAILED PROPOSAL SUBMISSION
70%
Literature Review
100%
Methodology Preparation
60%
Pre-existing Dataset Assessment
Technical Assessment
75%
DATASET COLLECTION AND PREPROCESSING
0%
Dataset Collection
0%
Dataset Cleaning
0%
Dataset Preprocessing
0%
ASR MODEL FINETUNING
0%
Model Implementation
0%
Testing and Evaluation
0%
Model Improvement and Fusion
0%
DOCUMENTATION
37%

Timeline - [2]


MAJOR PROJECT PART B
11-Aug-24
30-Sep-24
19-Nov-24
8-Jan-25
27-Feb-25
18-Apr-25
TRANSLATION MODEL FINETUNING
Model Implementation
Testing and Evaluation
Model Improvement and Fusion
SPEAKER ENCODER DEVELOPMENT
Model Implementation
Testing and Evaluation
Model Improvement and Fusion
MODIFYING FAST SPEECH
Model Implementation
Testing and Evaluation
Model Improvement and Fusion
MODEL CASCADING
Model Fusion
Testing and Evaluation
WEB INTERFACE
OVERALL TESTING
DOCUMENTATION
0%

Estimated Budget

S.N.	Items	Price(Rs.)
1.	Printing Expenses	5,000
2.	Fantech Leviosa Live MCX02 Microphone	6,000
3.	Miscellaneous	1,000
	Total	**12,000**

References - [1]

- Y. Wang et al., “Tacotron: Towards end-to-end speech synthesis,” arXiv.org, Mar. 29, 2017. https://arxiv.org/abs/1703.10135 (accessed Jun. 02, 2024).
- J. Shen et al., “Natural TTS synthesis by conditioning wavenet on mel spectrogram predictions,” arXiv.org, Dec. 16, 2017. https://arxiv.org/abs/1712.05884 (accessed Jun. 02, 2024).
- O. Kjartansson, S. Sarin, K. Pipatsrisawat, M. Jansche, and L. Ha, “Crowd-Sourced speech corpora for javanese, sundanese, sinhala, nepali, and bangladeshi bengali,” in 6th Workshop on Spoken Language Technologies for Under-Resourced Languages (SLTU 2018), Aug. 2018. Accessed: Jun. 17, 2024. [Online]. Available: http://dx.doi.org/10.21437/sltu.2018-11.

References - [2]

- K. Sodimana et al., “A Step-by-Step Process for Building TTS Voices Using Open Source Data and Frameworks for Bangla, Javanese, Khmer, Nepali, Sinhala, and Sundanese,” in 6th Workshop on Spoken Language Technologies for Under-Resourced Languages (SLTU 2018), Aug. 2018. Accessed: Jun. 17, 2024. [Online]. Available: http://dx.doi.org/10.21437/sltu.2018-14.
- Y. Ren et al., “FastSpeech 2: Fast and high-quality end-to-end text to speech,” arXiv.org, Jun. 08, 2020. https://arxiv.org/abs/2006.04558 (accessed Jun. 02, 2024).
- K. Zhou, B. Sisman, R. Liu, and H. Li, “Seen and Unseen Emotional Style Transfer for Voice Conversion with A New Emotional Speech Dataset,” in ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Jun. 2021. Accessed: Jun. 17, 2024. [Online]. Available: http://dx.doi.org/10.1109/icassp39728.2021.9413391.