Nepali To English Speech Translation With Prosody Prediction

Team Members

Pragyan Bhattarai (THA077BEI030)

Prashant Raj Bista (THA077BEI032)

Shakshi Kejriwal (THA077BEI044)

Sudipti Upreti (THA077BEI045)

Supervised By

Er. Kshetraphal Bohara

Department of Electronics and Computer Engineering
IOE, Thapathali Campus

August , 2024

Presentation Outline

- Introduction
- Motivation
- Problem statement and Objective
- Limitations
- Application
- Methodology
- Results and Analysis
- Discussion and Conclusion
- References

Introduction

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

Motivation

Problem Statement and Objectives

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 translated language.

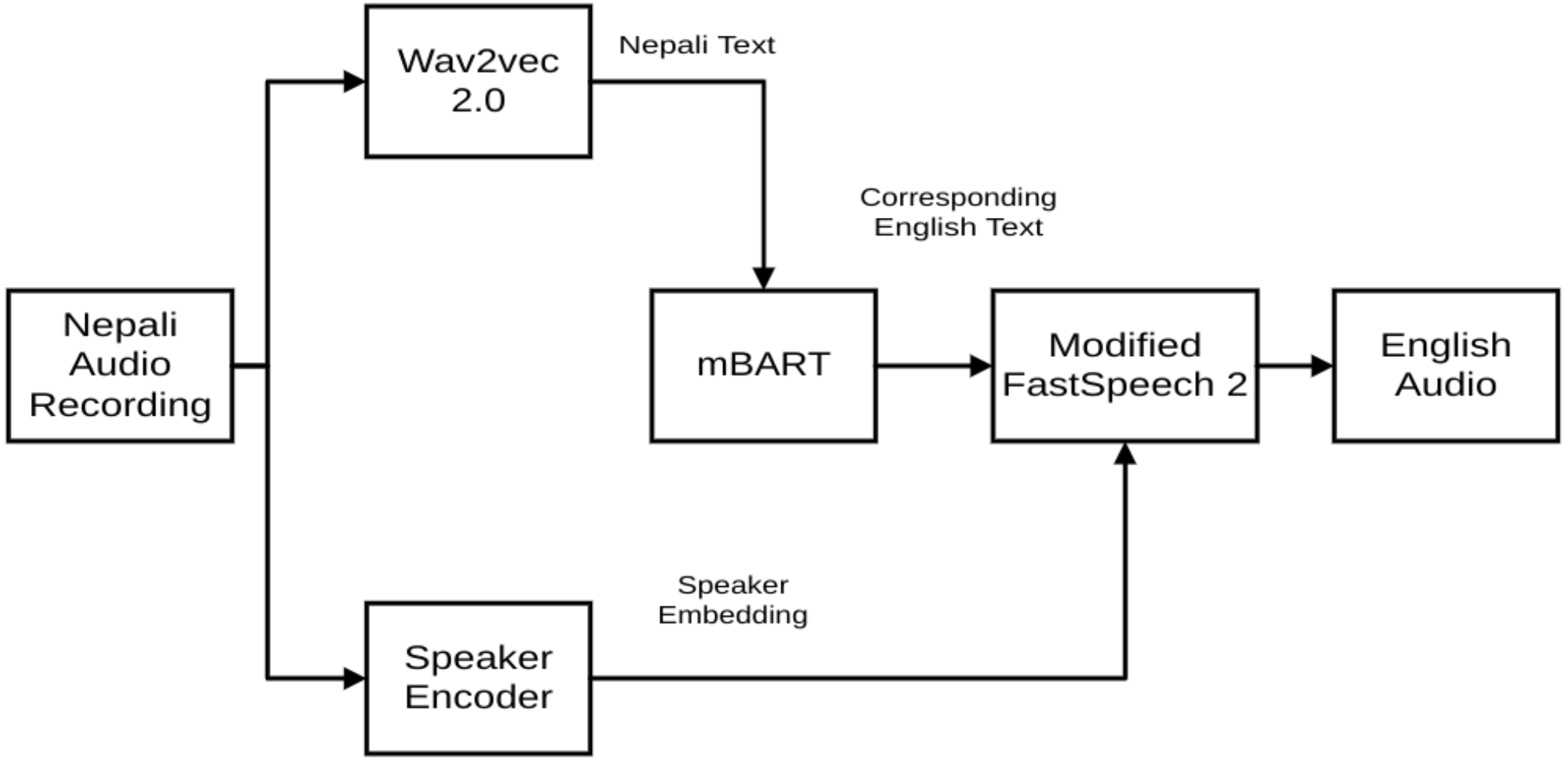
Limitations

- 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

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

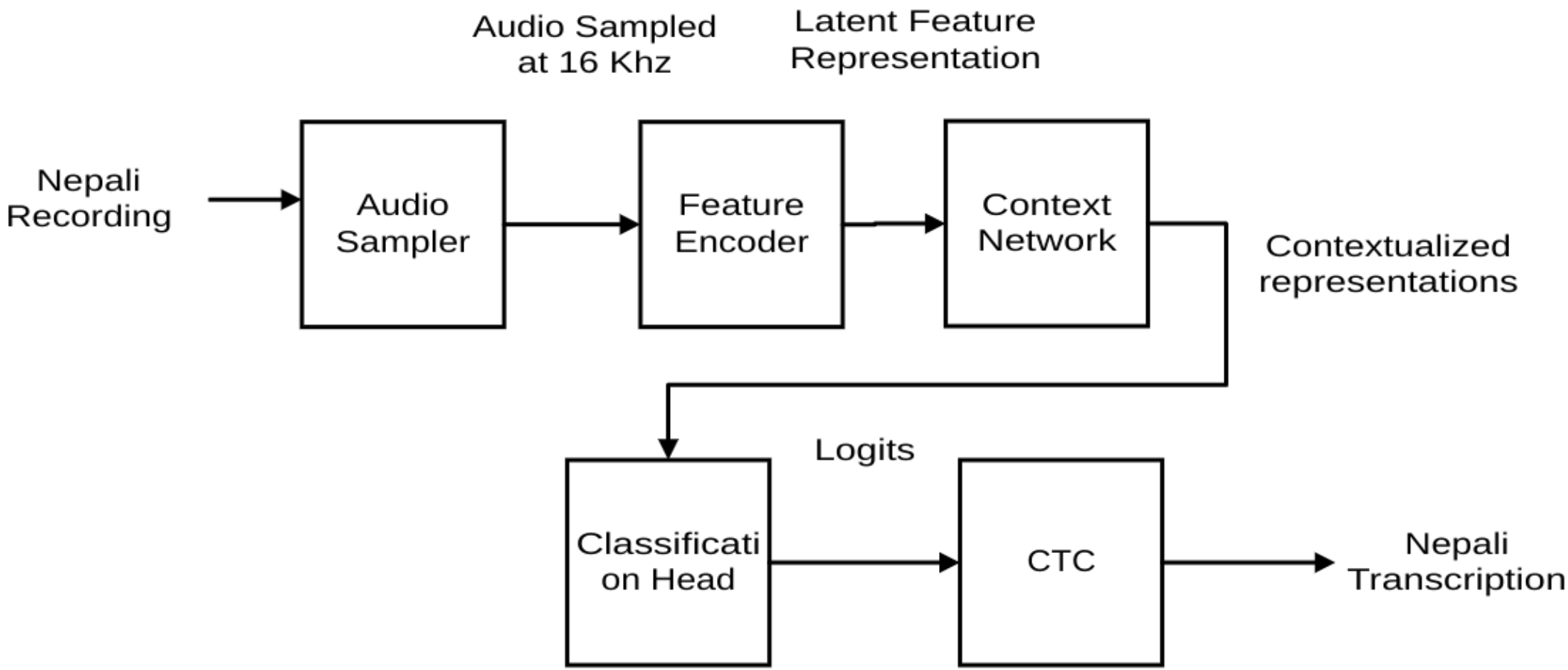
Methodology- [1] (Overall System Architecture)



Methodology - [2]
(Description of System Block Diagram)

- wav2vec 2.0 processes the recording.
- wav2vec 2.0 produces the corresponding Nepali text as output.
- mBART processes the Nepali text and outputs with an English Text.
- FastSpeech 2 takes English text and speaker embedding
- FastSpeech 2 synthesizes the corresponding English audio with desired prosody.

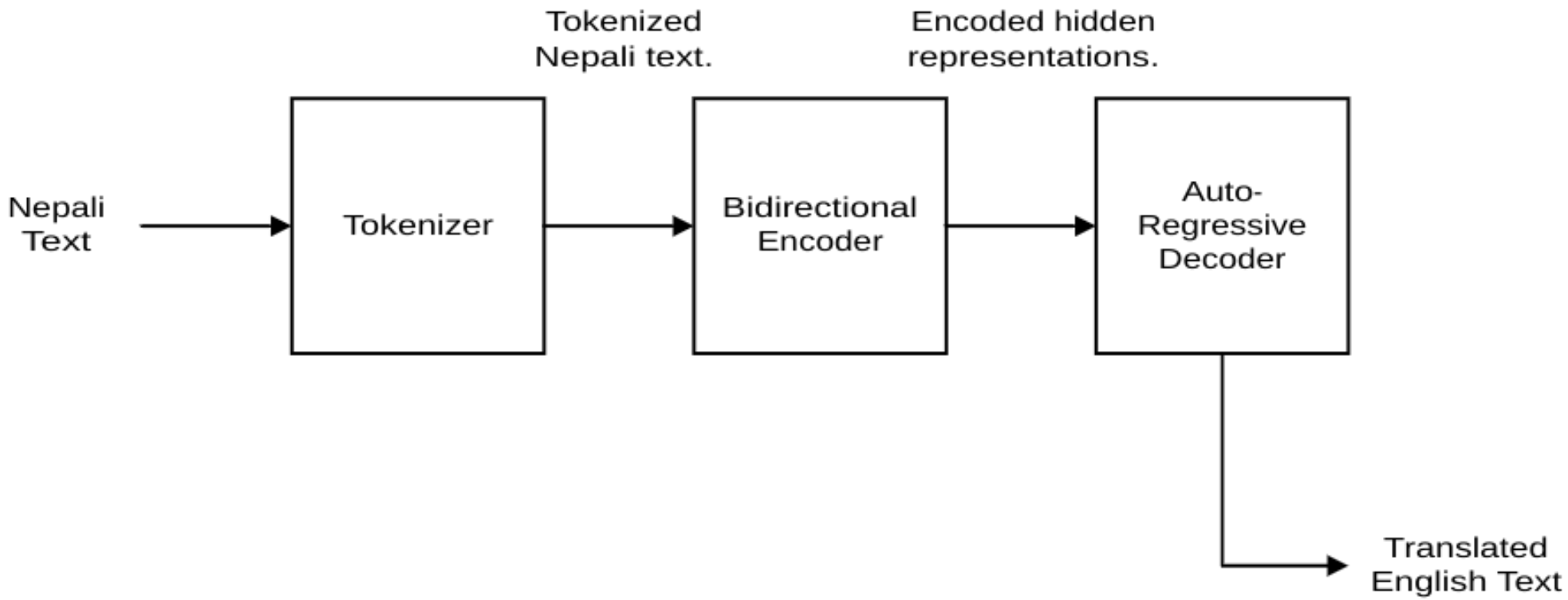
Methodology - [3] (Data Flow in wav2vec 2.0)



Methodology - [4] (wav2vec 2.0 Description)

- Feature encoder extracts the features and converts the data into the latent feature representation from recording sampled at 16KHz.
- Context network converts the latent feature into contextual representation.
- Classification head converts contextualized representation into the logits.
- CTC converts the logits into the corresponding Nepali text spoken in audio.

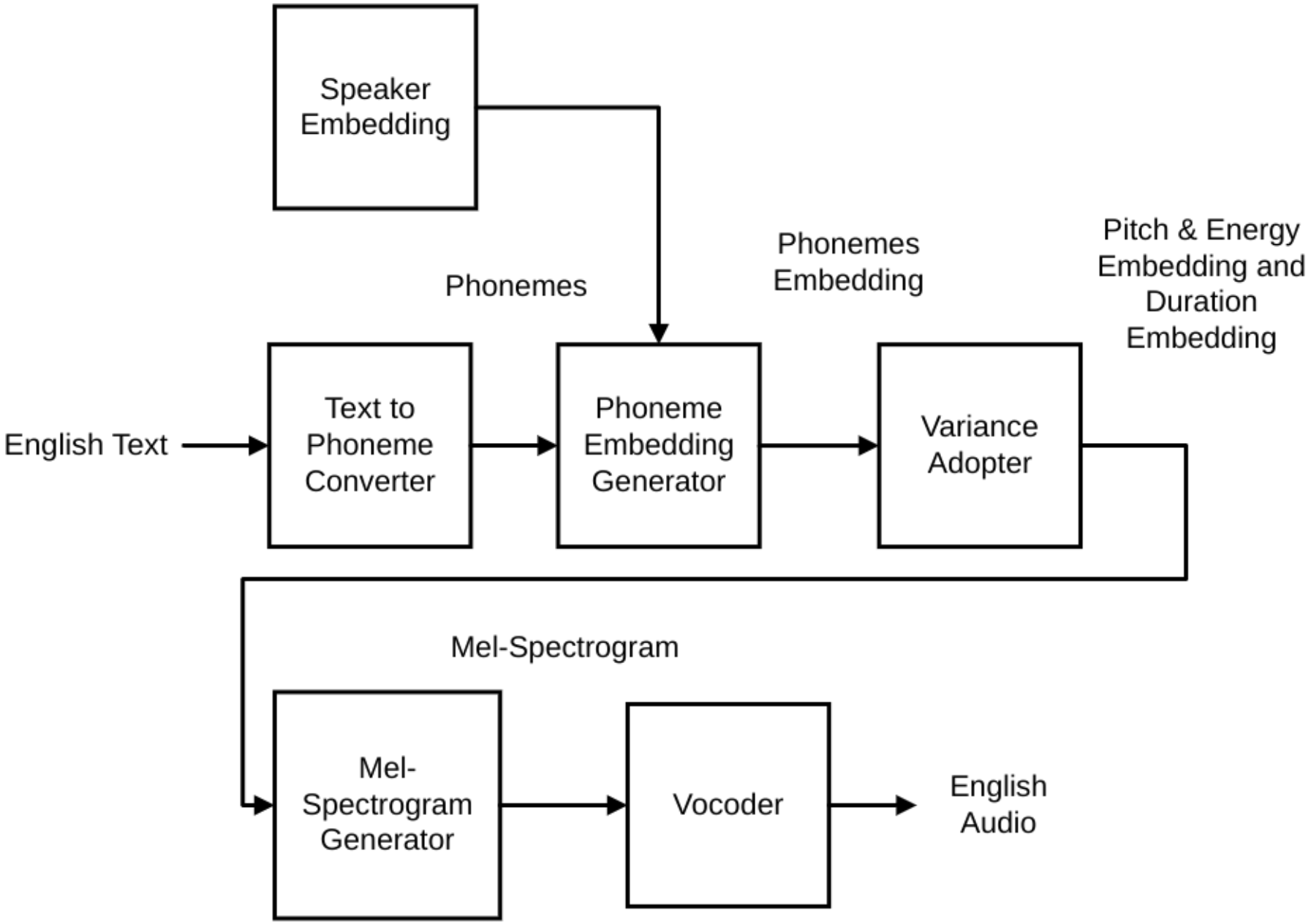
Methodology - [5] (mBART)



Methodology - [6] (mBART Description)

- Tokenizer converts the Nepali text into the tokens.
- Bidirectional encoder that produces the hidden representation of the data.
- Encoded hidden representation is passed to the autoregressive decoder.
- Decoder produces the transcribed English text.
- The transcribed English text is passed to a TTS model.

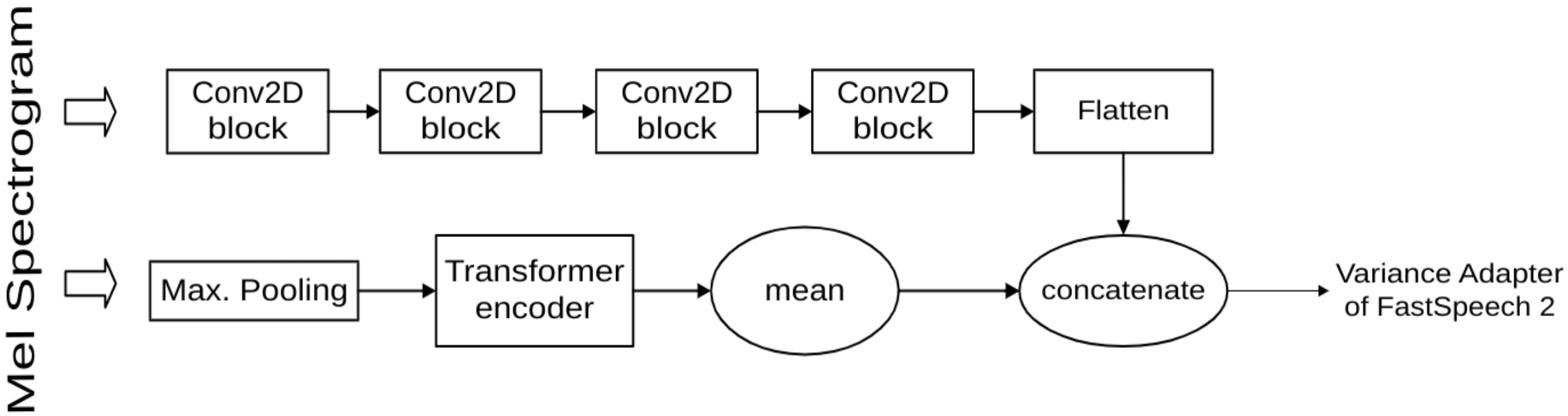
Methodology - [7] (FastSpeech 2)



Methodology - [8] (FastSpeech 2 Description)

- English text is converted to the respective phonemes.
- Phonemes encoder that produces the phoneme embeddings.
- Variance adopter predicts parameters for the generation of the speech.
- Speaker Embedding is concatenated with the output of the variance.
- Mel-Spectrogram generator generates the spectrogram for the speech to be synthesized.

Methodology - [9] (Speaker Encoder)



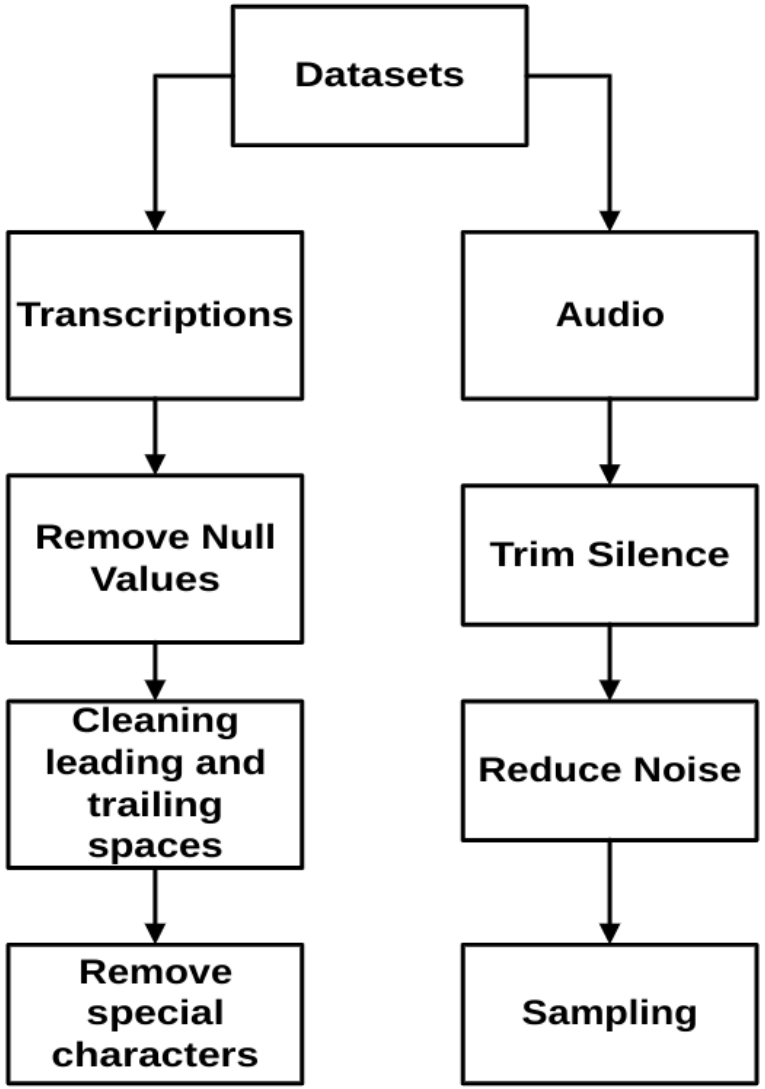
Methodology - [10] (Speaker Encoder Description)

- 2D convolution blocks captures hierarchical features from the input
- The output of the final Conv2D block is flattened into a 1D vector.
- max pooling is applied to reduce the dimensionality of the input features from Mel spectrogram.
- These pooled features are fed to Transformer encoder to capture dependencies in the data.

Methodology - [11] (Speaker Encoder Description)

- Output of the transformer encoder is averaged which summarizes the temporal features.
- Output of Conv2D and Transformer are concatenated forming single feature vector.
- Concatenation combines the local features catured by Conv2D blocks with the global features captured by Transformer encoder.
- The combined feature vector is fed into the Variance Adapter of TTS.

Methodology - [12] (Data Preprocessing Pipeline)



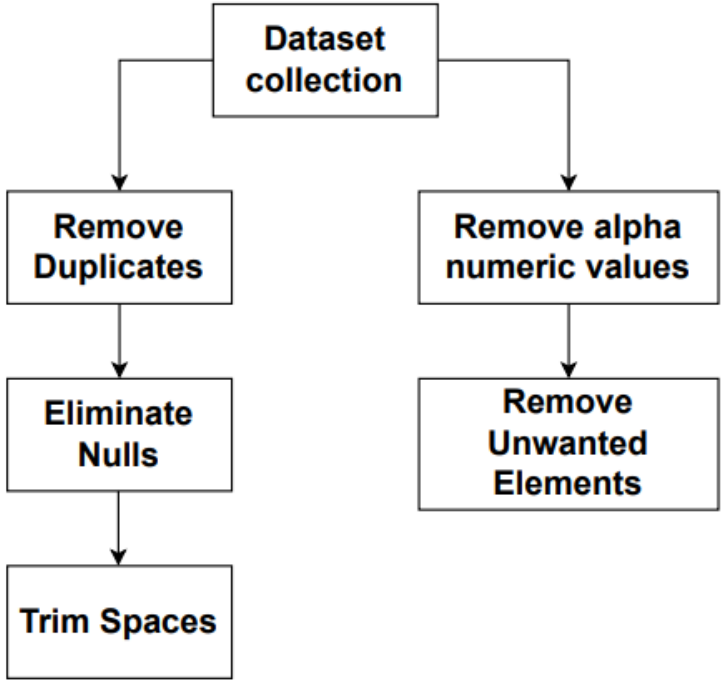
Source	Number of Audio
OpenSLR43	2064
OpenSLR143	675
Common Voice	772
Total	3511

Average Audio length: 4.5 seconds

Methodology - [13]
(Data Preprocessing Description)

- Datasets divided into transcriptions and audio.
- Transcriptions involve removing null values, cleaning leading and trailing spaces, and removing special characters.
- Audio includes trimming silence, reducing noise, and sampling.
- Ensures the data is clean and ready for further processing or analysis.

Methodology - [14]
(Data Preprocessing Pipeline mBART)

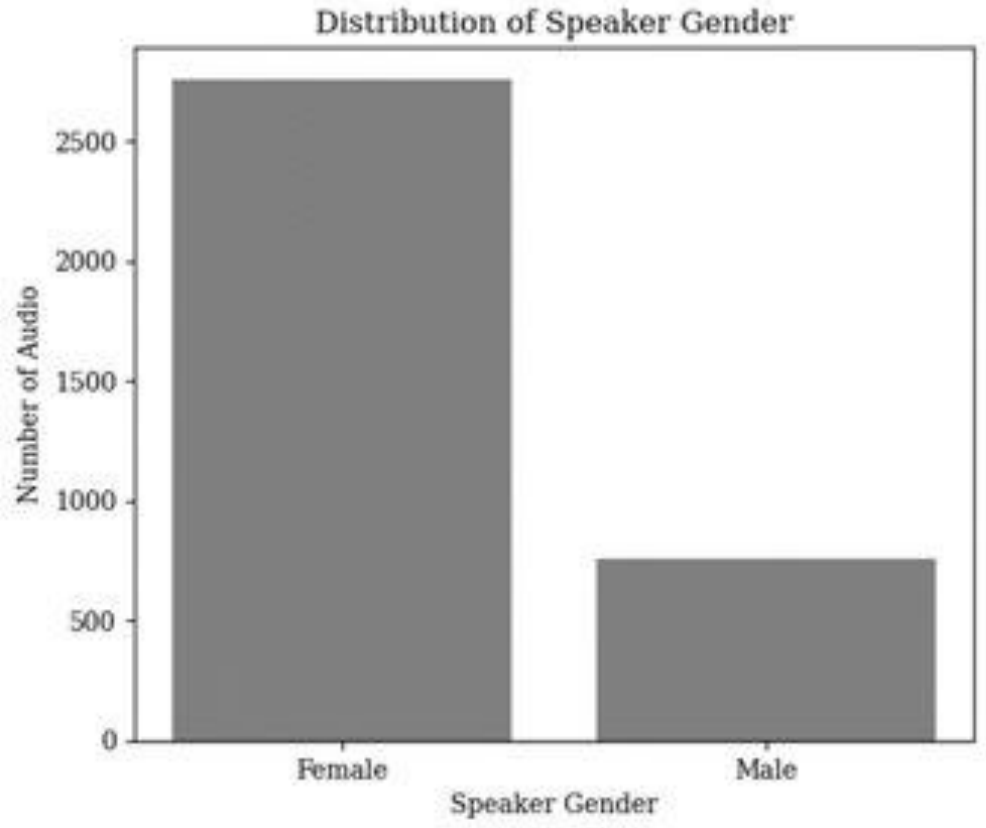


Corpus	Nepali Tokens	English Token
Wikimedia	289,460	280,771
TED2020	53,992	61,451
CCAligned	5,817,551	6,157,685
ParaCrawl	2,254,124	2,941,030
Bible-uedin	1,334,973	1,524,208
tico-19	63,371	70,587

Methodology - [15]
(Data Preprocessing Description mBART)

- Remove duplicates and null values to ensure data uniqueness and completeness.
- Trim spaces and convert text to lowercase to standardize the text format.
- Eliminate stopwords, symbols, numbers, punctuations, and URLs to clean unnecessary elements.

Result and Analysis - [1]
(ASR - Dataset Exploration)



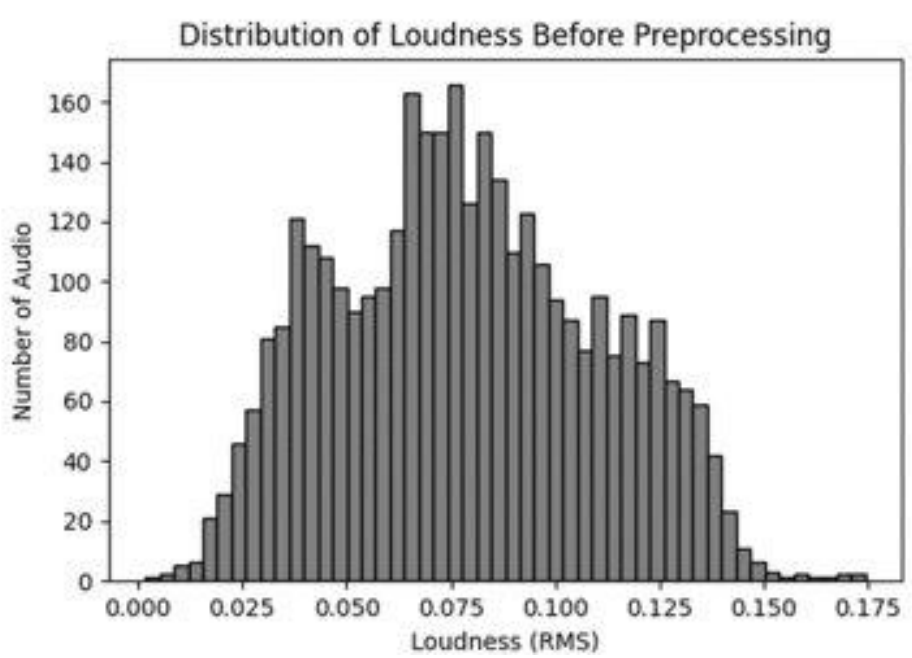
	Sentence	Length
Longest	म्याडम क्युरी फलोरेन्स नाइटिङ्गेल जुनको ताबेइ आदि यस्ता व्यक्तित्व हुन् जसले आफ्नो प्रतिभाको माध्य मबाट आफूलाई विश्वभर चिनाउन स मर्थ भए र यिनी सबै नारी नै हुन् जसले शिक्षा पाएका थिए	179
Shortest	सेवा	4

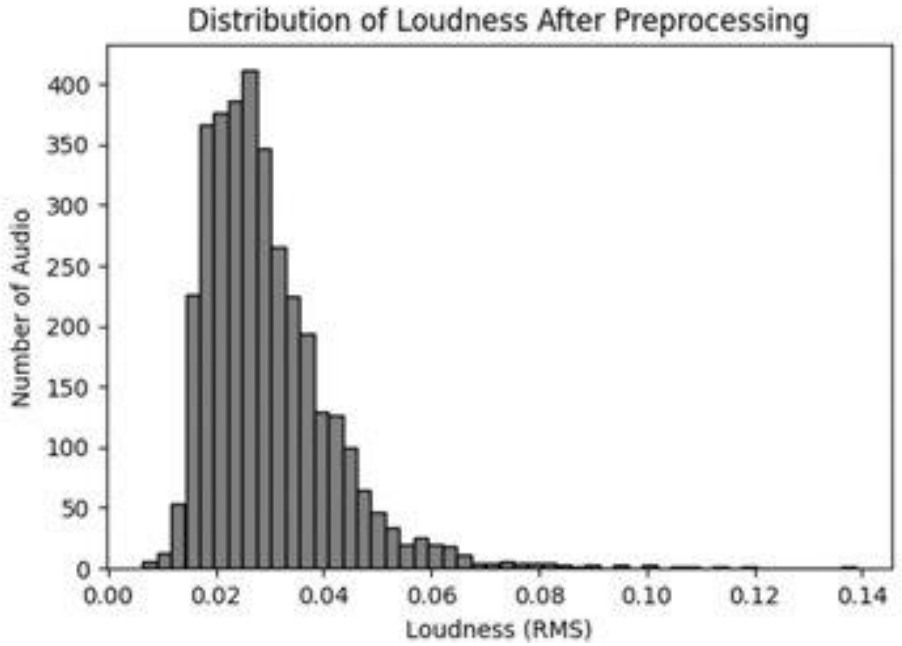
Result and Analysis - [2]
(ASR - Frequent and Rare Words)

Word	Frequency	Word	Frequnecy
छ	495	यो	178
हो	450	एक	162
पनि	263	छन्	142
रहेको	220	सय	132
भए को	207	जिल्ला मा	132

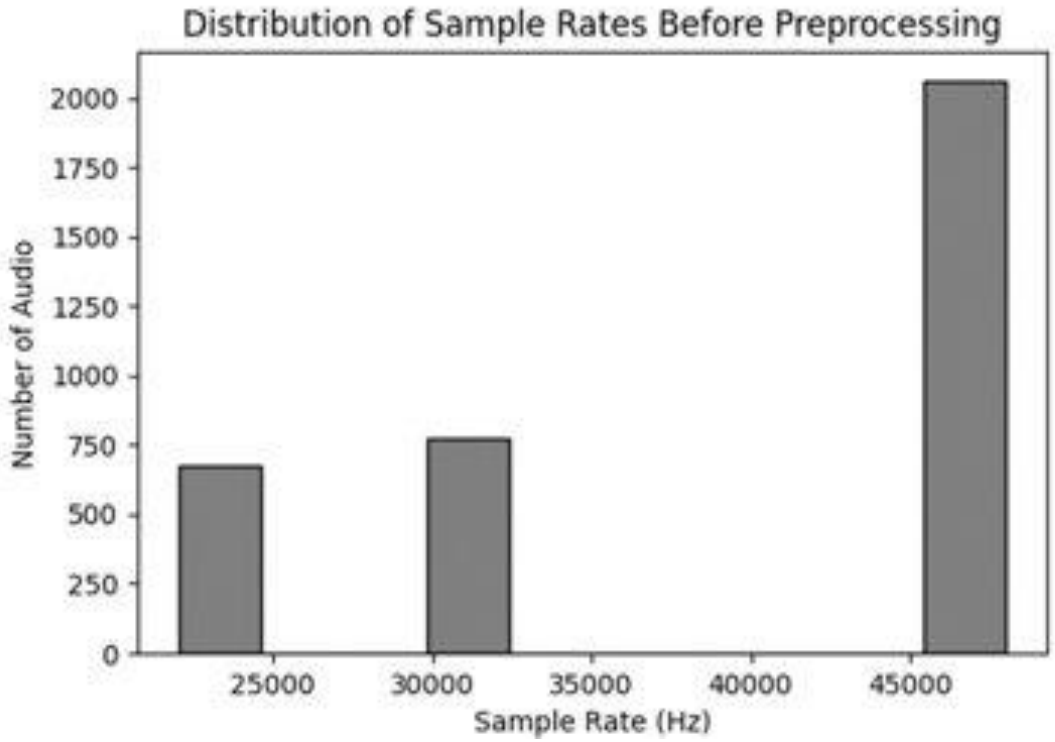
Word	Frequency	Word
आइरह न्छ	1	डुङ्गा
झट्का		जाऔ
भुकम्प को		लगाको
बगाउन		ओहो
भान्छा		कुर्र्सी

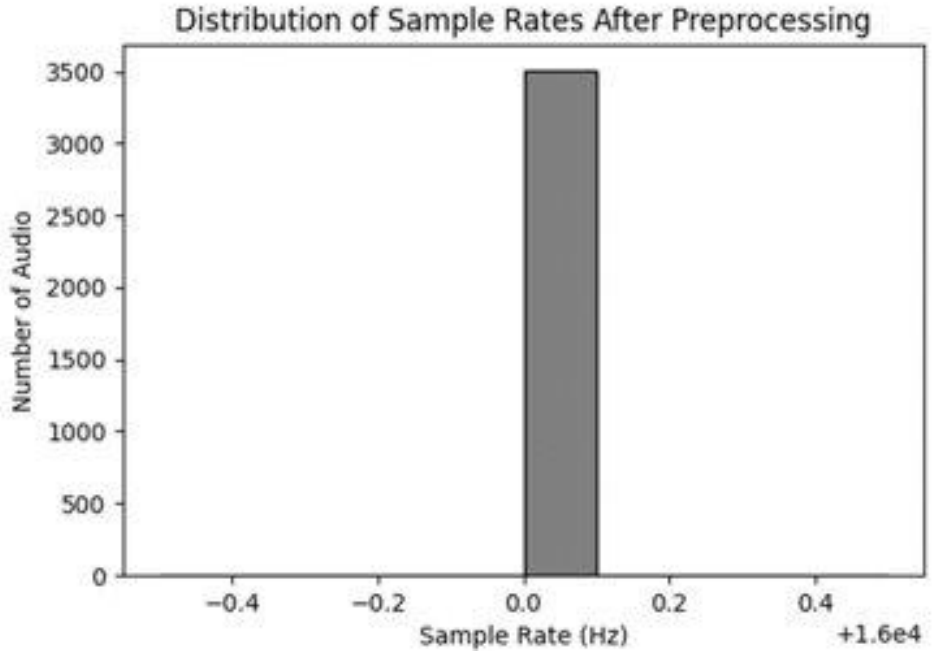
Result and Analysis - [3]
(Loudness Before and After Pre-Processing)



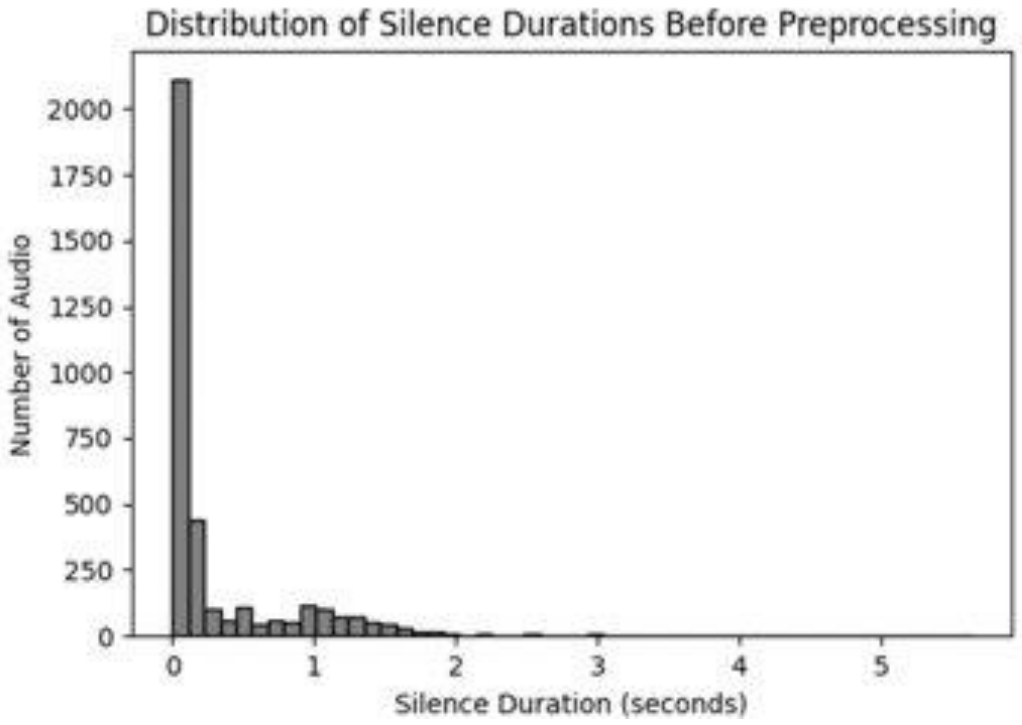


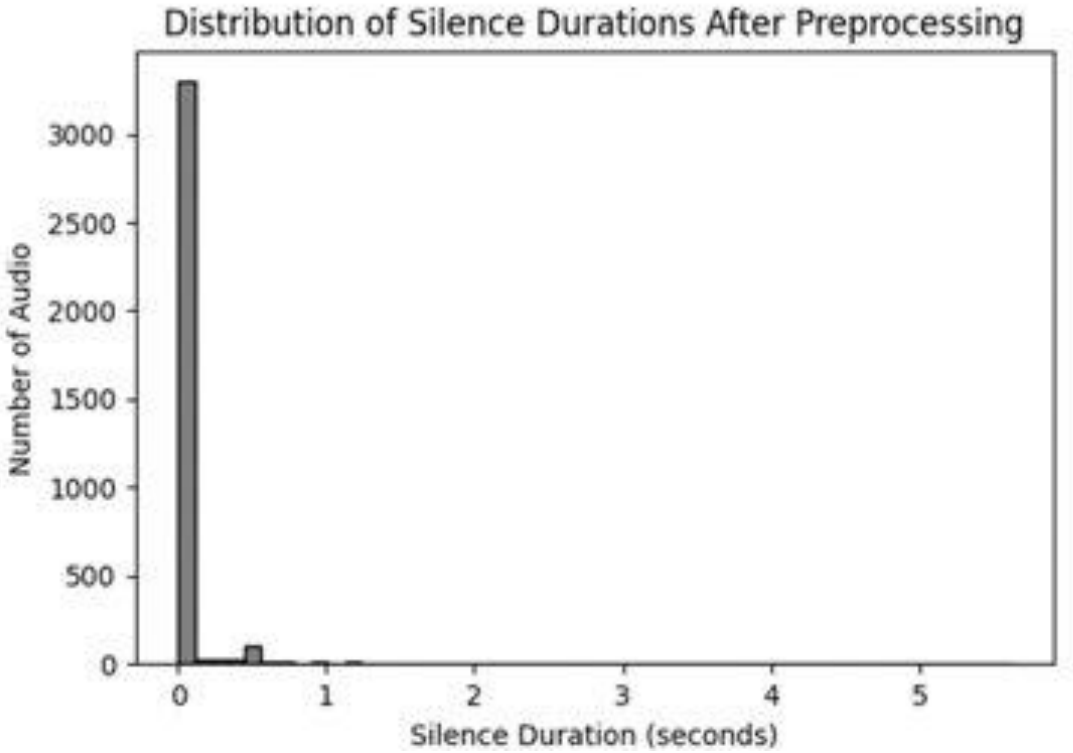
Result and Analysis - [4]
(Sampling Rate Before and After Pre-Processing)



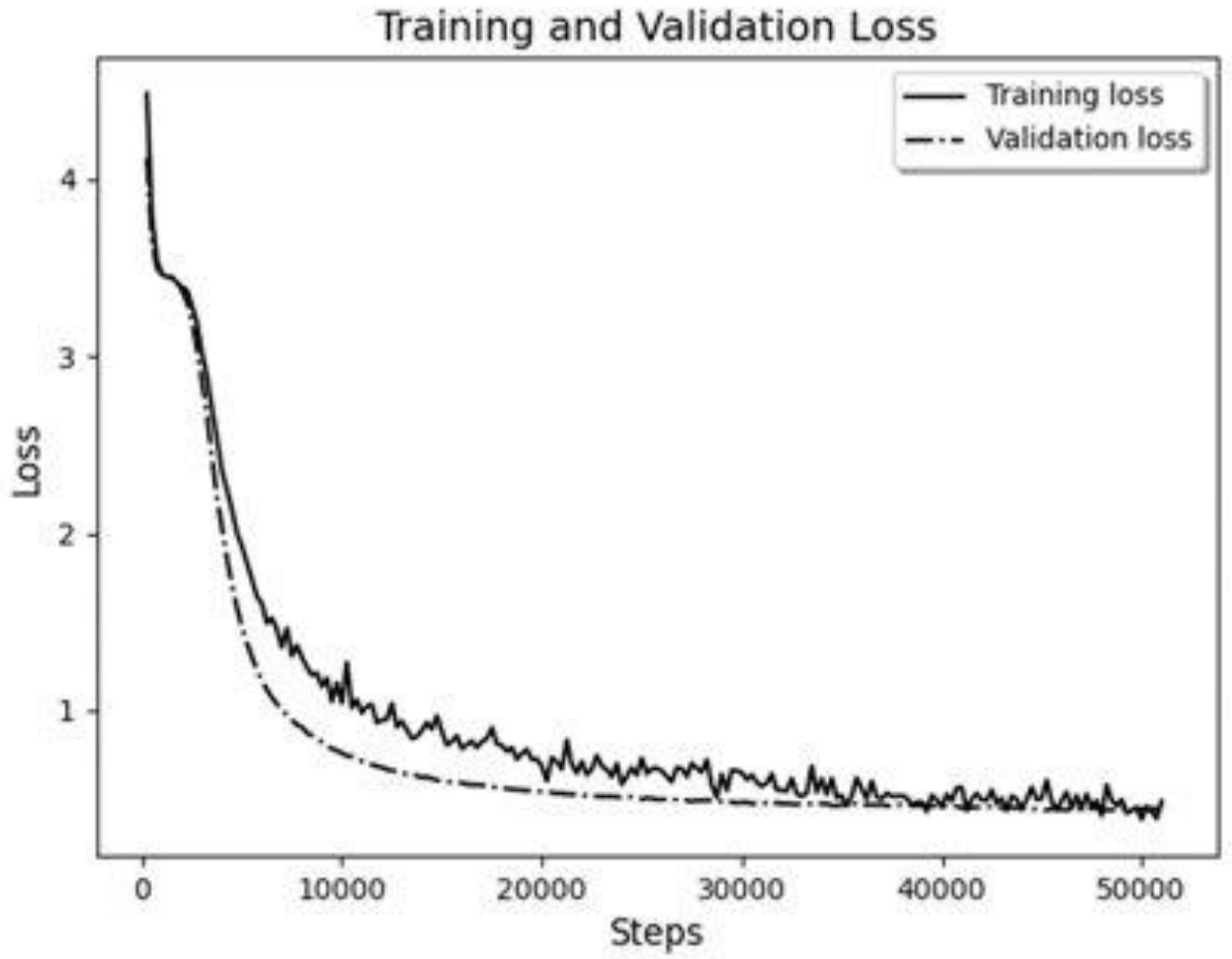


Result and Analysis - [5]
(Silence Duration Before and After Pre-Processing)

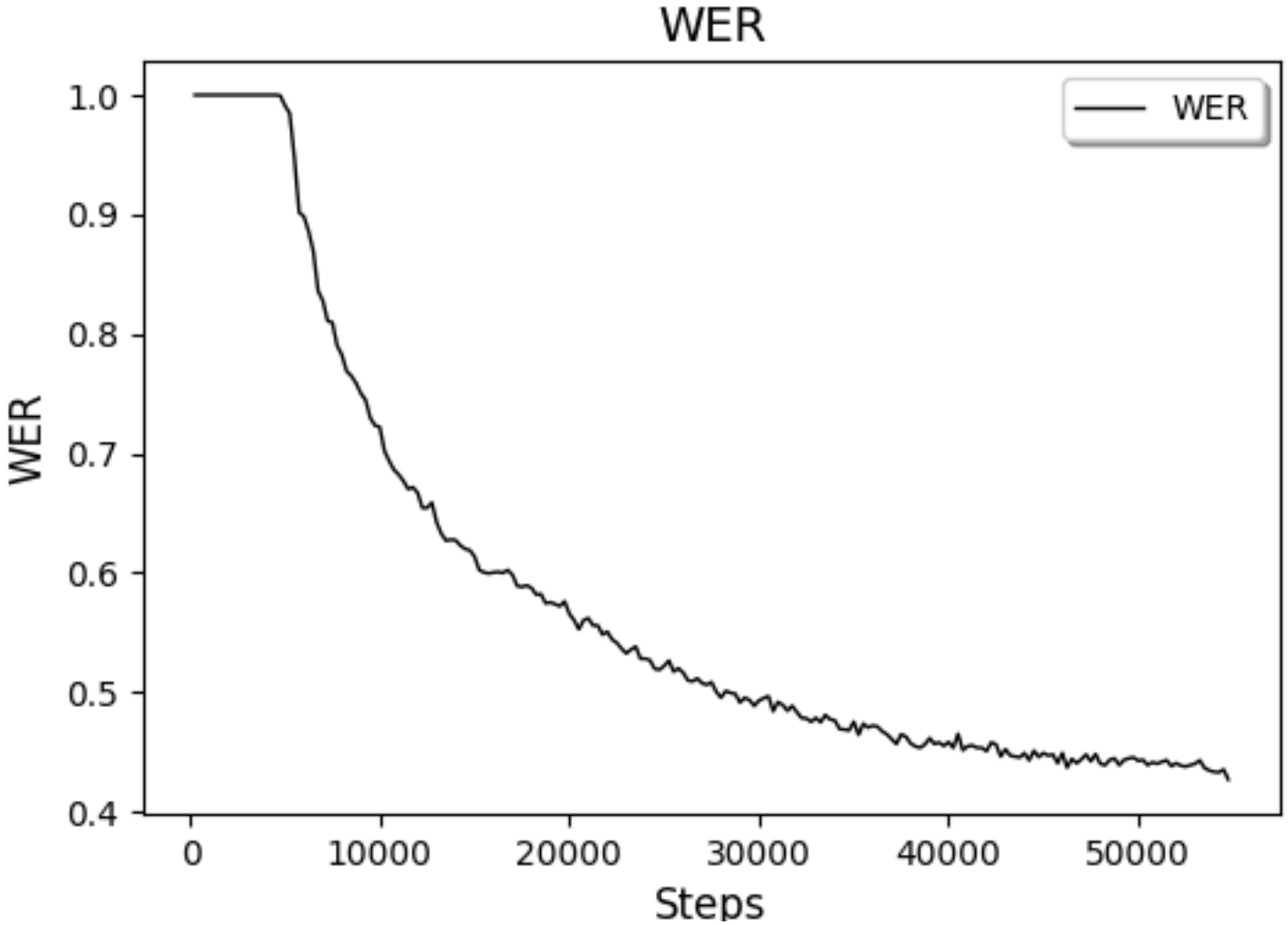




Result and Analysis - [6] (Training and Validation Loss of ASR)



Result and Analysis - [7]
(WER vs Steps curve for wav2vec 2.0)



Result and Analysis - [8] (Predictions from ASR)

True	Predicted	Count
र	हो	150
छ	र	139
र	छ	138
हो	र	131
छ	हो	123
हो	छ	113
पनि	र	79
पनि	हो	73
र	पनि	71
पनि	छ	70

Validation Samples: 281
Validation WER: 0.44

Test Samples: 703
Test WER: 0.4188

True	Predicted	Count
जाेशीकाे	जासीकाे	2
सूची	सुची	
कुत्री	नि	
भयो	भय	
यी	ी	
नजिकै	नजीकै	
बढी	बढि	
सम्पूर्ण	सम्पर्ण	

Result and Analysis – [9]
(Frequent words in NMT Dataset)

Word	Frequency	Word	Frequency
shall	10185	yahweh	5650
lord	7868	son	4861
god	7528	day	4534
unto	7170	man	4397
said	7129	israel	4396

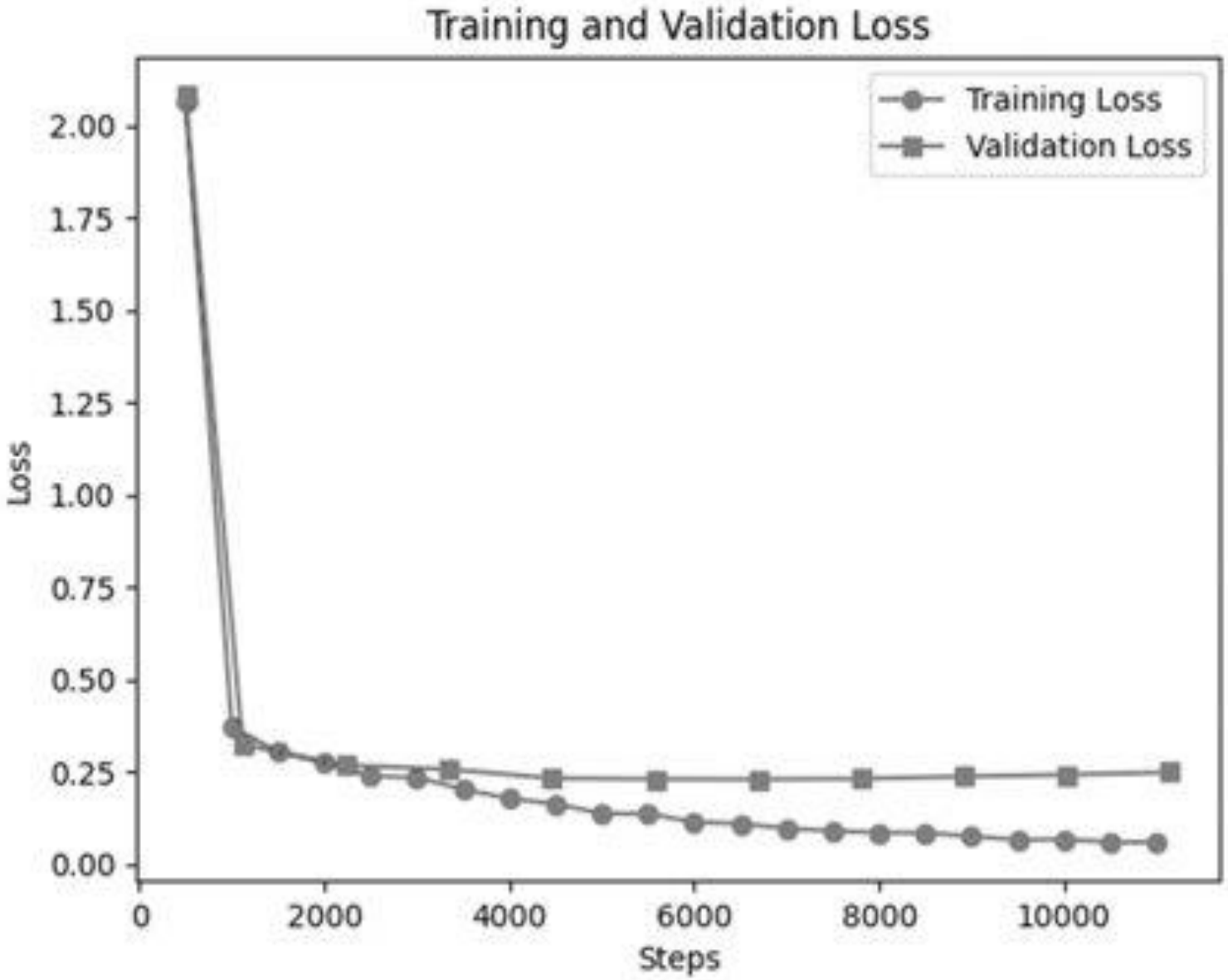
Word	Frequency	Word	Frequnecy
अनि	20829	परमेश्वर	9729
मानिसहरू	14340	तर	9629
तिनीहरू	13793	भने	8753
म	12982	तिमीहरू	7939
परमप्रभु	12177	त्यो	7726

Result and Analysis – [10]
(Rare Words in NMT Dataset)

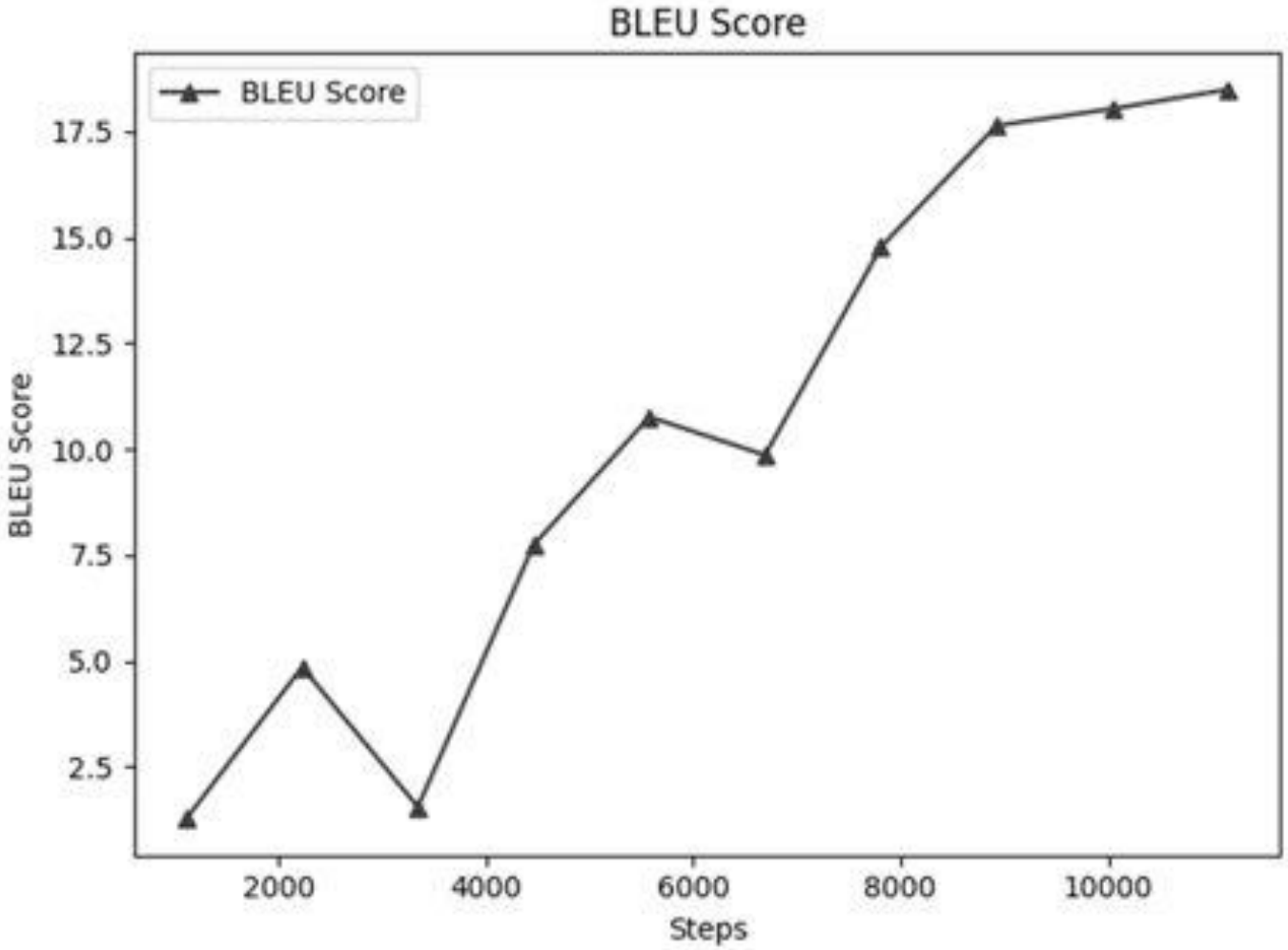
Word	Frequency	Word
quicker	1	immunisation
manufactured		nonvaccine
cytotoxicity		crs
antibody		refractory
neutralisation		retrospective

Word	Frequency	Word
जगाए	1	गैरखोप
सिन्ड्रोमहरू		सिआर्ईस
साइटोटक्सिसिटी		एन्टागोनिस्ट
एन्टिबडी		इन्टरल्युकिन
निष्प्रभावीकरण		रेट्रोस्पेक्टिभ

Result and Analysis - [11] (Training and Validaition Loss for mBART)



Result and Analysis - [12] (BLEU Score Curve for mBART)



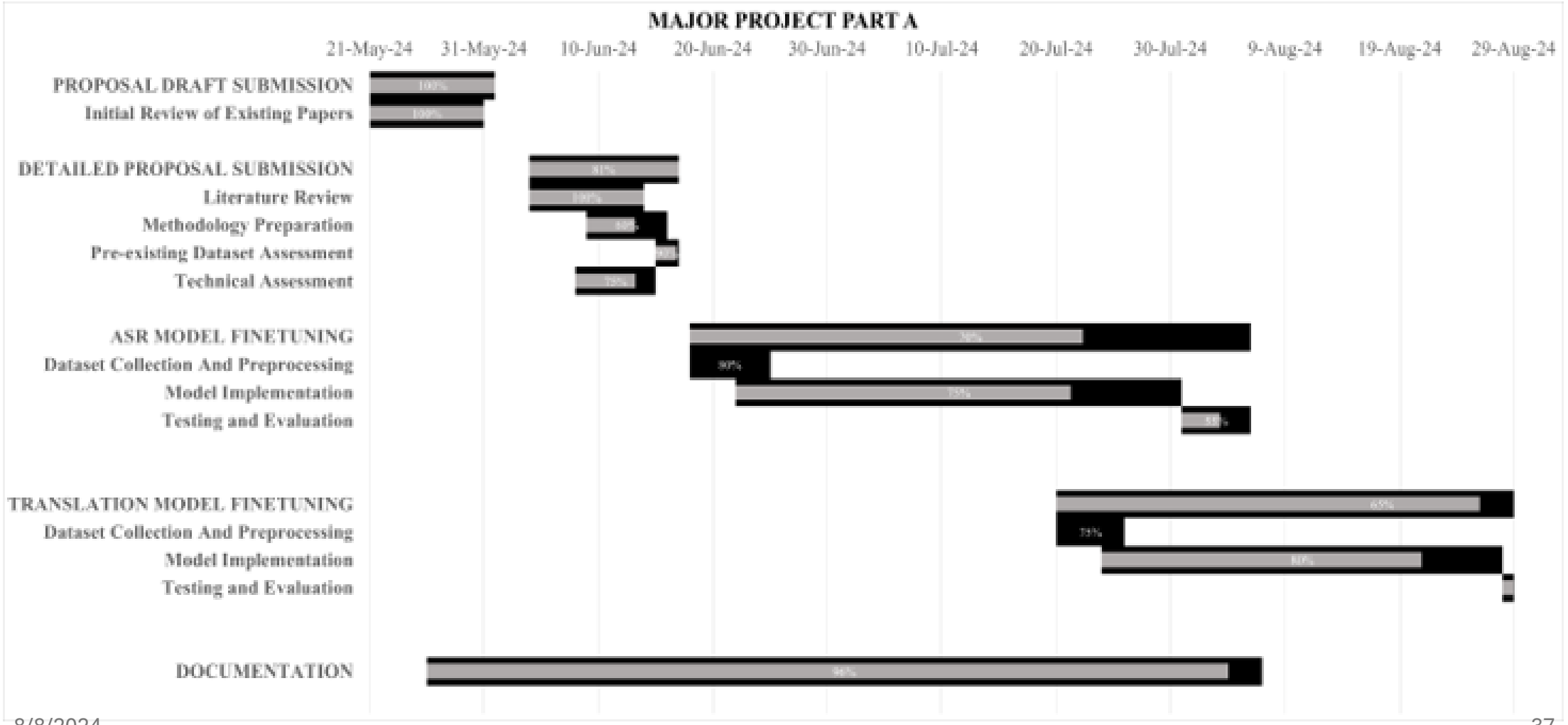
Conclusion

- WER for wav2vec 2.0 is 0.44.
  - Needs to be reduced by incorporating more datasets and experimenting with hyperparameters.
- BLEU for mBART is 18.48.
  - Indicates finetuning can improve Ne-En translation significantly compared to that of base model.
  - But, using 'Bible-uedin' dataset during training, caused the output to be inclined towards Old English.
  - Need for selection of better dataset for finetuning.

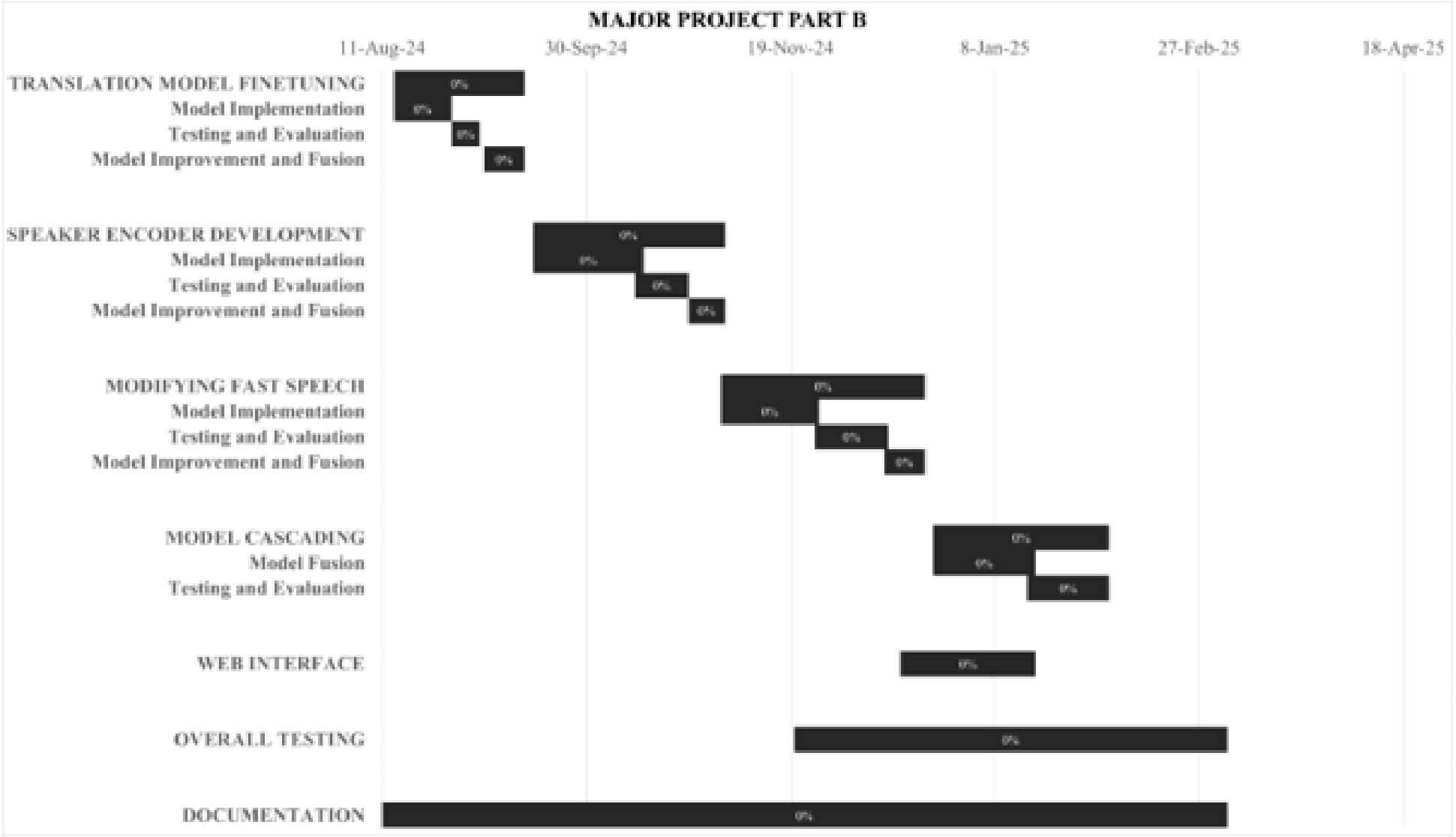
Remaining Task

- FastSpeech 2 Implementation
  - Develop and integrate the text encoder, duration, pitch, and energy predictors, and mel-spectrogram decoder.
- Model Training and Optimization
  - Train the model and optimize hyperparameters to improve performance
- Prosody Modelling
  - Extract prosodic features and integrate them into the FastSpeech 2 model.

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
Initial Review of Existing Papers
DETAILED PROPOSAL SUBMISSION
Literature Review
Methodology Preparation
Pre-existing Dataset Assessment
Technical Assessment
ASR MODEL FINETUNING
Dataset Collection And Preprocessing
Model Implementation
Testing and Evaluation
TRANSLATION MODEL FINETUNING
Dataset Collection And Preprocessing
Model Implementation
Testing and Evaluation
DOCUMENTATION

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

Budget

S.N.	Items	Price(Rs.)
1.	Printing expenses	5000
2.	Fantech Leviosa Live MCX02 Microphone	6000
3.	Miscellaneous	1000
	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.