Adapting Whisper AI for Multilingual Speech-To-Text Conversion in Bangladeshi Ethnic Languages

BY

Kabid Yeiad
ID: 202-15-14440

AND

Jannatul Ferdushi Jim
ID: 202-15-14439

FINAL YEAR DESIGN PROJECT REPORT

This Report Presented in Partial Fulfillment of the Requirements for the Degree of Bachelor of Science in Computer Science and Engineering

Supervised By

Ms. Sharun Akter Khushbu
Lecturer (Senior Scale)
Department of Computer Science and Engineering
Daffodil International University

Co-Supervised By

Mr. Abdus Sattar
Assistant Professor & Coordinator M.Sc
Department of Computer Science and Engineering
Daffodil International University

DAFFODIL INTERNATIONAL UNIVERSITY

DHAKA, BANGLADESH

July 2024

APPROVAL

This Project titled "**Adapting Whisper AI for Multilingual Speech-To-Text Conversion in Bangladeshi Ethnic Languages**", submitted by **Kabid Yeiad** and **Jannatul Ferdushi Jim** to the Department of Computer Science and Engineering, Daffodil International University, has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of B.Sc. in Computer Science and Engineering and approved as to its style and contents. The presentation has been held on 13th July.

BOARD OF EXAMINERS

Dr. Md. Ismail Jabiullah(MIJ) — **Board Chairman**
Professor
Department of CSE
Faculty of Science & Information Technology
Daffodil International University

Mr. Abdus Sattar (AS) — **Internal Examiner 1**
Assistant Professor
Department of CSE
Faculty of Science & Information Technology
Daffodil International University

Ms. Zahura Zaman (ZZ) — **Internal Examiner 2**
Lecturer
Department of CSE
Faculty of Science & Information Technology
Daffodil International University

Dr. Ahmed Wasif Reza (DWR) — **External Examiner**
Professor
Department of Computer Science and Engineering
East West University

DECLARATION

We hereby declare that this project has been done by us under the supervision of **Ms. Sharun Akter Khushbu, Lecturer (Senior Scale), Department of Computer Science and Engineering, Daffodil International University**. We also declare that neither this project nor any part of this project has been submitted elsewhere for the award of any degree or diploma.

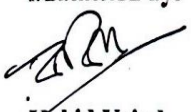
Supervised by:

Ms. Sharun Akter Khushbu
Lecturer (Senior Scale)
Department of CSE
Daffodil International University

Co-Supervised by:

Mr. Abdus Sattar
Assistant Professor & Coordinator M.Sc
Department of CSE
Daffodil International University

Submitted by:



Kabid Yeiad
ID: 202-15-14440
Department of CSE
Daffodil International University

Jannatul Ferdushi Jim
ID: 202-15-14439
Department of CSE

ACKNOWLEDGEMENT

First, we express our heartiest thanks and gratefulness to almighty for His divine blessing making it possible for us to complete the final year project/internship successfully.

We are grateful and wish our profound indebtedness to **Ms. Sharun Akter Khushbu**, Department of CSE Daffodil International University, Dhaka. Deep Knowledge & keen interest of our supervisor in the field of *NLP* to carry out this project. Her endless patience, scholarly guidance, continual encouragement, constant and energetic supervision, constructive criticism, valuable advice, reading many inferior drafts, and correcting them at all stages have made it possible to complete this project.

We would like to express our heartiest gratitude to the **Head of the Department of CSE**, for his kind help in finishing our project and also to other faculty members and the staff of the Department of CSE, Daffodil International University.

We express our sincere gratitude to **World Vision Bangladesh** for their invaluable assistance and unwavering support in facilitating the data collection process.

We would like to thank our entire course mate in Daffodil International University, who took part in this discussion while completing the course work.

Finally, we must acknowledge with due respect the constant support and patience of our parents.

ABSTRACT

This research shows how a full speech-to-text and translation system was made to turn sounds from five different ethnic languages into Bangla. These languages are Malo, Bawm, Santali, Marma, and Garo. The study looks at how well different neural network models, such as Bi-LSTM, CNN Seq2Seq, GRU Seq2Seq, Seq2Seq, and Transformer models, handle difficult jobs like translating and transcribing languages. To get the data, 20 native speakers of each language had to record 500 unique words. This made a strong sample that could be used for training and testing. Noise reduction, normalization, and segmentation were some of the preprocessing steps that made sure the inputs were of good quality. Adding more data made the model more stable. In real life, the Transformer model did better than others, with a 97% success rate in training and a 96% success rate in confirmation. The GRU Seq2Seq model also did well, finding a good balance between accuracy and speed. However, the CNN Seq2Seq model had a hard time. With high accuracy and low word error rates across all languages, the Whisper model did great at transcription jobs. The thorough test that used measures like accuracy, loss, validation accuracy, and validation loss showed that the Transformer model was the best at capturing long-range dependencies and context, which made it the best for translation. The Whisper model's reliable performance in transcription jobs is shown by how well it always does. This method has a big effect on people's lives because it lets them talk to each other in their native languages, brings people together, and protects cultural heritage. It gives minority language users more power by giving them access to basic services in their own languages. This makes it easier for them to be a part of society and improves their quality of life. This study builds a strong base for future improvements in speech-to-text and translation tools, which will help more people use languages and keep their cultures alive.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ACRONYMS

AI	Artificial Intelligence
ASR	Automatic Speech Recognition
BLEU	Bilingual Evaluation Understudy
Bi-LSTM	Bidirectional Long Short-Term Memory
CNN	Convolutional Neural Network
CSE	Computer Science and Engineering
GRU	Gated Recurrent Unit
IRJET	International Research Journal of Engineering and Technology
LSTM	Long Short-Term Memory
RNN	Recurrent Neural Network
Seq2Seq	Sequence-to-Sequence
S2T	Speech-to-Text
STT	Speech-to-Text
WER	Word Error Rate
AI	Artificial Intelligence
ASR	Automatic Speech Recognition

CHAPTER 1

Introduction

1.1 Overview

In the past few years, speech-to-text (STT) systems have made a lot of progress. This is because people need better and faster ways to communicate. There is a chance that these systems could help close language gaps, protect languages that are in danger of dying out, and make knowledge easier for people from all walks of life to access. Rapid progress in machine learning and artificial intelligence has made it possible to build complex models that can accurately transcribe and translate speech. But most of the study and development that has been done so far has been on mainstream languages. This means that ethnic and less-resourced languages have a big support gap. Languages from different ethnicities, like Bawm, Marma, Santali, Garo, and Malo, are important parts of culture and identity. Even though these languages are important, they don't always have the linguistic tools and technological support they need to be preserved and made available online. The complicated phonetics and syntax of these languages make it hard for current speech-to-text systems and translation models to work well. This means they aren't very useful in real life. New studies have found a number of problems with the way STT systems work now. For starters, a lot of these systems are made for languages that have a lot of training data, but ethnic languages don't have many labelled datasets, which makes it harder to make accurate models [1, 2]. Second, most STT models are trained on datasets that don't properly show the phonetic diversity and unique linguistic traits of ethnic languages [3]. This means that they don't work well with these languages. Third, there isn't a lot of study that specifically looks at how to test and improve STT models for ethnic languages. This means that we don't fully understand the best ways to use these models in these situations [6].

There isn't enough study on the need for effective speech-to-text and translation systems for ethnic languages right now. There is a big need for models that can correctly transcribe and translate speech from ethnic languages to languages that a lot of people speak, like Bangla. This gap shows how much more study needs to be done to make strong STT systems that are tailored to the specific needs of ethnic languages. The reason for this project is to help ethnic languages get access to technology so that they can be kept alive and used in the digital age.

The main purpose of this study is to create a complete speech-to-text system that accurately translates five ethnic languages to Bangla. These languages are Bawm, Marma, Santali, Garo, and Malo. In order to do this, the following steps will be taken: First, voice recordings from people who speak the five target ethnic languages will be used to make a big dataset. Twenty different speakers will record in each language, and each person will add 500 unique words. Second, several advanced neural network

models, such as Transformer, Seq2Seq, GRU Seq2Seq, CNN Seq2Seq, Bi-LSTM will be created and tested. The collected information will be used to train these models and find the best way to do transcription and translation Third, a full comparison report of the models will be made to find the model that works the best. The test will focus on how well the person can be accurate, how quickly they can work, and how well they can deal with the unique aspects of each ethnic language. Lastly, the model that was picked because it is the most accurate and stable will be put into a real-world speech-to-text system. This method will be put to the test in the real world to make sure it works and is easy to use. The successful completion of this research will not only contribute to the field of speech-to-text technology but also provide a valuable tool for preserving and promoting the use of ethnic languages. This study aims to set a new standard for STT systems in the area of ethnic language translation by filling in the gaps and solving the current problems. This will help make the digital world more diverse and open to everyone.

1.2 Background and Present State

Ethnic languages have a crucial role in preserving cultural history and individual identity, but they are becoming more endangered as a result of globalization and the prevalence of dominant languages. Ensuring the conservation and availability of these languages is essential for upholding cultural variety and historical coherence. The present speech-to-text (STT) technologies, propelled by breakthroughs in machine learning and artificial intelligence, have demonstrated substantial potential in converting spoken language into written text. However, these systems primarily cater to languages that have abundant linguistic databases and significant computational resources. In contrast, ethnic languages frequently face a shortage of annotated datasets and linguistic tools that are specifically designed to meet their unique requirements. The current condition of STT systems highlights a notable disparity: current models, including those utilising Whisper AI and diverse neural network structures, face difficulties in dealing with the intricate phonetic and syntactic aspects of ethnic languages, resulting in less than optimal performance [1][2]. Furthermore, there is a significant dearth of targeted research on the development and enhancement of STT systems for these languages, which worsens the difficulties associated with digital preservation and accessibility. Although attempts have been made to tackle these problems, such as developing small voice databases and modifying general models for low-resource languages, these endeavors generally lack accuracy and resilience [4]. This study aims to fill these knowledge deficiencies by creating a comprehensive Speech-to-Text (STT) system capable of precisely transcribing and translating five ethnic languages (Bawm, Marma, Santali, Garo, Malo) into Bangla. This will be achieved by utilising advanced neural network models to improve linguistic inclusion and preserve cultural heritage.

1.3 Problem Statement

The accurate translation of ethnic languages remains a significant challenge due to their unique phonetic structures and limited linguistic resources, despite the significant advancements in speech-to-text (STT) technology. The performance of current STT systems, which are predominantly designed for well-resourced languages, is inadequate when applied to ethnic languages such as Bawm, Marma, Santali, Garo, and Malo. The transcription and translation accuracy of these languages is suboptimal due to a scarcity of annotated datasets and tailored linguistic tools. The poor performance and limited practical utility of existing models are frequently the result of their inability to generalize across the diverse phonetic and syntactic features that are inherent in these languages. This disparity not only exacerbates the digital divide but also impedes the preservation and accessibility of ethnic languages. Consequently, there is a pressing necessity for a focused research endeavor to create a STT system that is both reliable and precise, capable of accurately translating ethnic languages into Bangla. This system should utilize sophisticated neural network models to promote cultural preservation and linguistic inclusivity. The objective of this research is to resolve these obstacles by generating an exhaustive dataset, assessing numerous neural network models, and implementing the most effective model to offer a practical solution for ethnic language translation.

1.4 Objectives

The main goal of this project is to create a precise and effective speech-to-text (STT) system that can accurately translate five ethnic languages—Bawm, Marma, Santali, Garo, and Malo—into Bangla. In order to accomplish this, the study intends to gather an extensive dataset of voice recordings from individuals who are native speakers of different languages. Each participant will provide 500 distinct words. The research will entail the creation and assessment of several sophisticated neural network models, such as Transformer, Seq2Seq, GRU Seq2Seq, CNN Seq2Seq, Bi-LSTM, and Character Based Seq2Seq, in order to determine the best efficient model for this particular application. The project aims to identify the most effective technique by conducting a comprehensive comparative analysis of these models. This analysis will consider factors like as accuracy, efficiency, and the models' capacity to handle the distinct phonetic and syntactic characteristics of each ethnic language. The primary objective is to integrate the selected model into a functional STT system, which can then be evaluated and verified in real-life situations. This will help in safeguarding and making ethnic languages more accessible, hence promoting linguistic inclusiveness.

1.5 Scope and Limitation

The main goal of the research is to create a speech-to-text (STT) system that can translate between Bangla and five different regional languages: Bawm, Marma, Santali, Garo, and Malo. As part of the study, a large dataset of voice recordings from native speakers will be gathered, with each speaker adding 500 unique words. To find the best model for this purpose, the study is creating and testing a number of advanced neural network models, such as Transformer, Seq2Seq, GRU Seq2Seq, CNN Seq2Seq, Bi-LSTM, and Character Based Seq2Seq. The study's goal is to find the model that works best in terms of accuracy, speed, and being able to handle the unique phonetic and syntactic features of each ethnic tongue. For the final application, the chosen model will be tested and proven to work in real life to make sure it is useful. The Limitations are:

- ·The research is limited to the translation of the selected five ethnic languages to Bangla and does not extend to other languages or dialects.
- ·The dataset, while comprehensive, may not capture all possible dialectal variations and speech patterns within each ethnic language, which could affect the model's generalizability.
- ·The performance of the STT system may vary based on the recording environment, speaker accents, and background noise, which were not uniformly controlled during data collection.
- Due to resource constraints, extensive field testing and further refinements of the models may be limited, potentially impacting the robustness and scalability of the final system.
- The study focuses on a single target language (Bangla) for translation, which may not fully address the multilingual needs of users who speak other languages in addition to Bangla.

1.6 Report Organization

The intent of this research is to create a speech-to-text (STT) system that can translate five ethnic languages—Bawm, Marma, Santali, Garo, and Malo—into Bangla. The study's scope encompasses the acquisition of a comprehensive dataset that includes voice recordings from native speakers, with each speaker contributing 500 unique words. The research entails the development and assessment of a variety of sophisticated neural network models, such as Transformer, Seq2Seq, GRU Seq2Seq, CNN Seq2Seq, Bi-LSTM, and Character Based Seq2Seq, in order to determine the most effective model for this application. The objective of the investigation is to identify the model that exhibits the highest level of accuracy, efficiency, and the capacity to accommodate the distinctive phonetic and syntactic characteristics of each ethnic language.

The final implementation will entail the testing and validation of the selected model in real-world scenarios to guarantee its practical utility. The research is restricted to the transformation of the five ethnic languages that have been chosen to Bangla and does not encompass any other languages or dialects. The model's generalizability may be impacted by the fact that the dataset, despite its comprehensiveness, may not encompass all possible dialectal variations and speech patterns within each ethnic language. The STT system's efficacy may be contingent upon the recording environment, speaker accents, and background noise, which were not consistently managed during data collection. Extensive field testing and additional model refinements may be restricted as a result of resource constraints, which could potentially affect the final system's scalability and robustness. The research concentrates on a single target language (Bangla) for translation, which may not adequately satisfy the multilingual requirements of users who speak other languages in addition to Bangla.

1.7 Summary

This research focuses on the urgent requirement for a proficient speech-to-text (STT) system that can accurately convert five ethnic languages—Bawm, Marma, Santali, Garo, and Malo—into Bangla. Although there have been notable improvements in speech-to-text (STT) technology, current models are not able to effectively handle the distinct linguistic characteristics and restricted resources of ethnic languages. As a result, these models have poor performance and accessibility concerns. The objective of this study is to fill this void by creating a strong collection of voice recordings from individuals who speak the language well and assessing various sophisticated neural network models to determine the most efficient method. The selected model will be applied and evaluated in practical situations, thereby aiding in the conservation and availability of ethnic languages and promoting linguistic inclusiveness. This research aims to overcome these problems and establish a higher benchmark for translating ethnic languages. Additionally, it aims to contribute to the digital conservation of cultural assets.

CHAPTER 2

LITERATURE REVIEW

2.1 Overview

This chapter presents a thorough examination of the literature about speech-to-text (STT) systems, specifically addressing the difficulties and remedies for converting ethnic languages, like Bangla, into widely spoken languages. The purpose of the literature review is to provide a comprehensive understanding of the current status of STT technology, emphasize the shortcomings and deficiencies in previous research, and pinpoint the unresolved matters that this study attempts to tackle. More specifically, we will examine previous research that has made significant contributions to the area of Speech-to-Text (STT), with a particular focus on studies that address the challenges of working with languages that have limited resources and efforts to preserve ethnic languages. We will analyses and juxtapose these works in order to comprehend their approaches, merits, and limitations. This comparison will assist in identifying the specific areas where current models are deficient, particularly in their ability to handle the distinct phonetic and syntactic characteristics of ethnic languages. In addition, we will analyses the unresolved concerns in the field, such as the limited availability of annotated datasets, the challenges in adapting existing models to other scenarios, and the requirement for realistic implementations that can be verified in real-world situations. This chapter will provide the basic information necessary for constructing a strong and efficient speech-to-text (STT) system specifically designed for translating ethnic languages. It will thoroughly analyses these factors, paving the way for the following chapters on methodology, results, and debate.

2.2 Related Works

Significant progress has been made in the development of speech-to-text (STT) systems, with a primary emphasis on languages that have ample resources. Wang et al. [1] presented Fairseq S2T, a framework that prioritises rapidity and effectiveness in neural network-based speech-to-text modelling. Although it achieves remarkable outcomes, it fails to tackle the distinct difficulties presented by ethnic languages, which frequently suffer from a scarcity of comprehensive information. In a similar vein, Nagdewani and Jain [2] give a thorough examination of speech-to-text (STT) and text-to-speech (TTS) conversion techniques. They point out a lack of research specifically focused on ethnic languages, which are not adequately represented in mainstream STT studies. Hasan et al. [3] enhanced the precision of transcribing Bangla by incorporating a spell-checker into a speech-to-text (STT) system. Although this demonstrates that additional tools can improve particular languages, it fails to acknowledge the intricacies

of various ethnic languages that have unique phonetic and syntactic characteristics. Ahmed et al. [4] developed a Bangla speech corpus using audio and text materials that were publically accessible. This corpus was established to aid in the development of speech-to-text (STT) technology. Nevertheless, this endeavour fails to address the limited availability of resources for ethnic languages or the necessity for multilingual speech-to-text systems. Zhang et al. [5] investigated the use of transfer learning methods to enhance speech conversion systems by utilising non-parallel training data. While this showcases the possibility of using models trained on languages with abundant resources to languages with limited resources, it primarily emphasises voice conversion rather than direct speech-to-text translation and does not explicitly address ethnic languages. Patil et al. [6] created a bilingual automatic speech recognition (ASR) model that utilises attention mechanisms. Their research demonstrates the potential of this model in efficiently processing voice input from two languages. Nevertheless, its potential to adapt and expand to accommodate many ethnic languages has not yet been evaluated. Bell et al. [7] conducted a review of adaptation methods for voice recognition systems based on neural networks. Their study offers valuable insights on enhancing model performance through adaptation. However, the review does not provide particular experimental findings for languages spoken by ethnic groups. Chen and Yang [8] utilised deep learning techniques to identify and understand speech in the Yi language, emphasising the capability to work with languages that have little resources. Nevertheless, this study is constrained to only one language and does not encompass multilingual situations or translation assignments. Dey and Dutta [9] introduced a compact automatic speech recognition (ASR) system specifically developed for edge devices with limited resources. The solution prioritizes efficiency and minimal computing demands. Although the system is well-suited for many situations, its accuracy and resilience may be constrained when it comes to complicated ethnic languages. Nugroho and Noersasongko [10] improved the accuracy of Indonesian ethnic speaker recognition by employing data augmentation, showcasing its efficacy. Nevertheless, their research is specifically centred on speaker identification rather than speech-to-text translation. Hasan and Islam [11] conducted a study on the identification of emotions in Bengali speech using Recurrent Neural Network (RNN) models. Although their approaches may be advantageous for the development of STT systems, they do not specifically tackle the issues associated with translating ethnic languages. Nayak et al. [12] investigated the use of deep learning for recognising spoken commands in the low-resource KUI language. Their study emphasised the practicality of employing deep learning techniques in languages with limited resources. However, its functionality is restricted to the recognition of spoken commands and does not encompass complete speech-to-text translation. Williams et al. [13] assessed the effectiveness of Wav2Vec2 and Whisper models in Automatic Speech Recognition (ASR) for the low-resource Maltese language. They emphasised that both models are suitable for low-resource languages, however their evaluation was limited to a single language and did not consider their performance in multilingual settings. Jain et al. [14]

modified Whisper models to enhance the accuracy of infant voice recognition. However, they did not specifically address the problems of translating ethnic languages.

Radhakrishnan et al. [15] proposed a cross-modal generative error correction framework for speech recognition, with the goal of enhancing mistake correction. However, their approach does not specifically focus on ethnic language translation. Kodali et al. [16] utilised Wav2Vec2 and Whisper embeddings to categorise vocal intensity levels. Their study specifically focused on vocal intensity classification for ethnic languages, rather than speech-to-text (STT) or translation tasks. Park et al. [17] suggested a methodology for improving the accuracy of speech-to-text (STT) systems by post-processing. This method involves utilising text-to-speech and back-transcription techniques. Nevertheless, the potential use of this approach in ethnic languages has not been investigated. Finally, Subakan et al. [18] utilised attention processes to enhance speech separation, showcasing their efficacy. However, their focus was mostly on speech separation rather than translation, and they did not expressly address ethnic languages.

2.3 Comparison between existing works

Table 1. Comparative analysis with previous work

SL No	Author Name	Used Algorithm	Best Accuracy with Algorithm
1.	Ott et al. [1]	RNN-based, Transformer-based, Conformer-based	—
2	Nagdewani, S., & Jain, A. (2020)[2]	Speech-To-Text (STT): Hidden Markov Model (HMM) with Deep Neural Network (DNN) Text-To-Speech (TTS): Hidden Markov Model (HMM)	**STT**: combination of HMM with DNN for STT conversion is considered the most suitable technique
3	Hasan, H. M., Islam, M. A., Hasan, M. T., Hasan, M. A., Rumman, S. I., &	DeepSpeech, CMUSphinx	DeepSpeech 96%

	Shakib, M. N. (2020, August)[3]		
4	Shafayat Ahmed, et al.	Hybrid CTC-Attention	20.2% WER
5	Mingyang Zhang, et al. [5]	Non-Parallel Voice Conversion	—
6	Aditya Patil et al.[6]	Bilingual MultiSoftmaxAttn	13.3% on Hindi, 8.23% on English, 1.3% on Code-mixed
7	Peter Bell, Joachim Fainberg, Ondrej Klejch, Jinyu Li, Steve Renals, Pawel Swietojanski[7]	Various Neural Network Adaptation Algorithms	—
8	Ziyan Chen, Hongwu Yang[8]	TDNN	
9	Swarnava Dey, Jeet Dutta[9]	RNN-T	WER: 18.65%
10	Nugroho, K., & Noersasongko, E.[10]	DA-DNN7L (Data Augmentation Deep Neural Network with 7 Layers)	99.76% Accuracy
11	H. M. Mahmudul Hasan, Md. Adnanul Islam[11]	RNN with Categorization (WC)	51.33%

12	Subrat Kumar Nayak et al. [12]	Attention using LSTM	97%
13	Aiden Williams, Andrea Demarco, Claudia Borg[13]	Wav2Vec 2.0 and Whisper	8.53% WER (CommonVoice), 24.98% WER (MASRI)
14	Jain, R., Barcovschi, A., Yiwere, M., Corcoran, P., & Cucu, H.[14]	Whisper, Wav2vec2 (finetuned)	—
15	S. Radhakrishnan, C. H. H. Yang, S. A. Khan, R. Kumar, N. A. Kiani, D. Gomez-Cabrero, J. N. Tegner [15]	Cross-modal Fusion for Error Correction	37.66% Improvement in WER Relative Performance
16	Kodali, M., Kadiri, S., & Alku, P. [16]	Wav2vec2, Whisper, SVM	—

2.4 Open Issues

Although there have been notable progressions in the domain of speech-to-text (STT) systems, there are still unresolved difficulties, specifically related to the conversion of ethnic languages. One of the most urgent obstacles is the lack of available data. Many ethnic languages do not have complete and annotated datasets, which are essential for training precise and resilient speech-to-text (STT) models. The absence of sufficient data impedes the capacity of models to acquire the distinct phonetic and syntactic characteristics of these languages, leading to subpar performance. Another notable concern pertains to the capacity of current models to generalize. The majority of speech-to-text (STT) systems are trained using datasets that do not sufficiently capture the wide range of phonetic variations and complex linguistic characteristics of ethnic

languages. As a result, these models frequently struggle to apply their knowledge effectively to languages spoken by specific ethnic groups, resulting in low accuracy when transcribing and unreliable translations.

The limited research focus on constructing models specifically customized to ethnic languages, as highlighted by Nagdewani and Jain [2], worsens this situation. The present speech-to-text (STT) technologies have limited ability to translate many languages. Although certain models, such the ones created by Patil et al. [6], have demonstrated potential in dealing with bilingual situations, their capacity to handle various ethnic languages on a wide scale has not been thoroughly investigated. Efficient translation models must possess the ability to handle the varied phonetic and grammatical structures of many languages, which is an intricate challenge that has not been sufficiently addressed in the current body of literature. Moreover, there is a significant dearth of practical applications and empirical verification for speech-to-text (STT) systems specifically developed for ethnic languages. Several investigations, such as the ones conducted by Chen and Yang [8] and Ahmed et al. [4], primarily concentrate on the construction of theoretical models and the evaluation of small-scale scenarios. Nevertheless, it is crucial to carry out actual implementation and thorough field testing to guarantee the models' usability and efficacy in real-life situations.

The lack of actual implementation hinders the effectiveness and availability of these technologies for the populations that require them the most. Furthermore, the exploration of integrating supplementary tools and employing post-processing techniques is still limited. Hasan et al. [3] showed the advantages of using a spell-checker to enhance accuracy. However, there is a lack of equivalent improvements for speech-to-text systems in ethnic languages. Additional methods such as context-aware correction, phonetic matching, and domain-specific adaptation have the potential to greatly improve the performance of speech-to-text (STT) systems for ethnic languages. The main unresolved problems in the domain of speech-to-text (STT) systems for ethnic languages encompass the scarcity of data, restricted ability of models to generalize, inadequate capacities for multilingual translation, absence of practical implementations, and insufficient integration of auxiliary tools. To tackle these issues, it is necessary to allocate focused research endeavors, create extensive datasets, and construct sophisticated models that are customised to the unique requirements of ethnic languages. By directing attention towards these specific domains, forthcoming investigations have the potential to greatly enhance the precision, effectiveness, and pragmatic use of speech-to-text (STT) systems for languages of various ethnicities. This, in turn, will promote a broader range of linguistic variety and inclusiveness in the realm of digital technologies.

2.5 Summary

This chapter has presented a thorough examination of the current body of literature pertaining to speech-to-text (STT) systems, specifically emphasizing ethnic languages. The evaluation emphasized notable progress in speech-to-text (STT) technologies, including the Fairseq S2T framework [1], as well as diverse approaches for converting voice to text and text to speech [2]. Notwithstanding these progressions, there are still several gaps and issues that have not been dealt with. The majority of previous research has concentrated on languages that have sufficient resources, while ethnic languages, which frequently lack comprehensive annotated datasets and have distinctive phonetic and syntactic difficulties, have received less consideration. The literature review showed that whereas auxiliary tools like spell-checkers can improve transcription accuracy for certain languages [3], there is a requirement for comprehensive solutions that address numerous ethnic languages. Research conducted by Patil et al. [6] and Nugroho and Noersasongko [10] demonstrate possible approaches for enhancing speech-to-text (STT) systems using bilingual models and data augmentation, respectively. However, the extent to which these approaches may be used to ethnic languages and translation tasks has not been thoroughly investigated. The reported open difficulties encompass data scarcity, restricted model generalisation, inadequate language translation capabilities, lack of practical implementations, and insufficient integration of auxiliary tools. To tackle these issues, it is necessary to allocate focused research endeavours, create extensive datasets, and design sophisticated models that are customized to the particular requirements of ethnic languages.

CHAPTER 3

METHODOLOGY

3.1 Overview

This chapter provides a full explanation of the approach used to create a strong system that can transcribe and translate audio from five different ethnic languages (Malo, Bawm, Santali, Marma, and Garo) into Bangla. The methodology includes various essential elements: a comprehensive data collection phase that involves recording the voices of native speakers to capture genuine linguistic characteristics, followed by meticulous data preprocessing steps such as noise reduction, normalization, and segmentation to guarantee the quality of the data. The primary aspect of the process entails the creation and assessment of various sophisticated neural network models, including Bi-LSTM, CNN-Bi-LSTM, CNN-Seq2Seq, GRU-Seq2Seq, Seq2Seq, and Transformer models. The objective is to determine the most efficient model for precise transcription and translation.

The Whisper model was picked for the transcription phase because of its resilience in dealing with various accents, while the Transformer model was selected for translation, showcasing exceptional performance with a training accuracy of 97% and a validation accuracy of 96%. In addition, data augmentation techniques were utilized to improve the resilience of the model by introducing fluctuations in the audio samples. The hardware infrastructure, consisting of top-of-the-line GPUs, SSD storage, and ample RAM, played a vital role in managing the computational requirements of model training and data processing.

The project was guided to successful completion by the use of effective project management, which included well-defined timetables and milestones, as well as rigorous financial analysis to assure adherence to the budget. The systematic and comprehensive strategy employed ensured the creation of a very precise and effective system for translating ethnic languages into Bangla, effectively tackling the distinct linguistic difficulties presented by these languages.

3.2 Proposed Methodology

The proposed methodology for developing the system to transcribe and translate audio from five ethnic languages—Malo, Bawm, Santali, Marma, and Garo—into Bangla involves several key stages: data collection and preprocessing, model development and evaluation, and the final integration of the transcription and translation models.

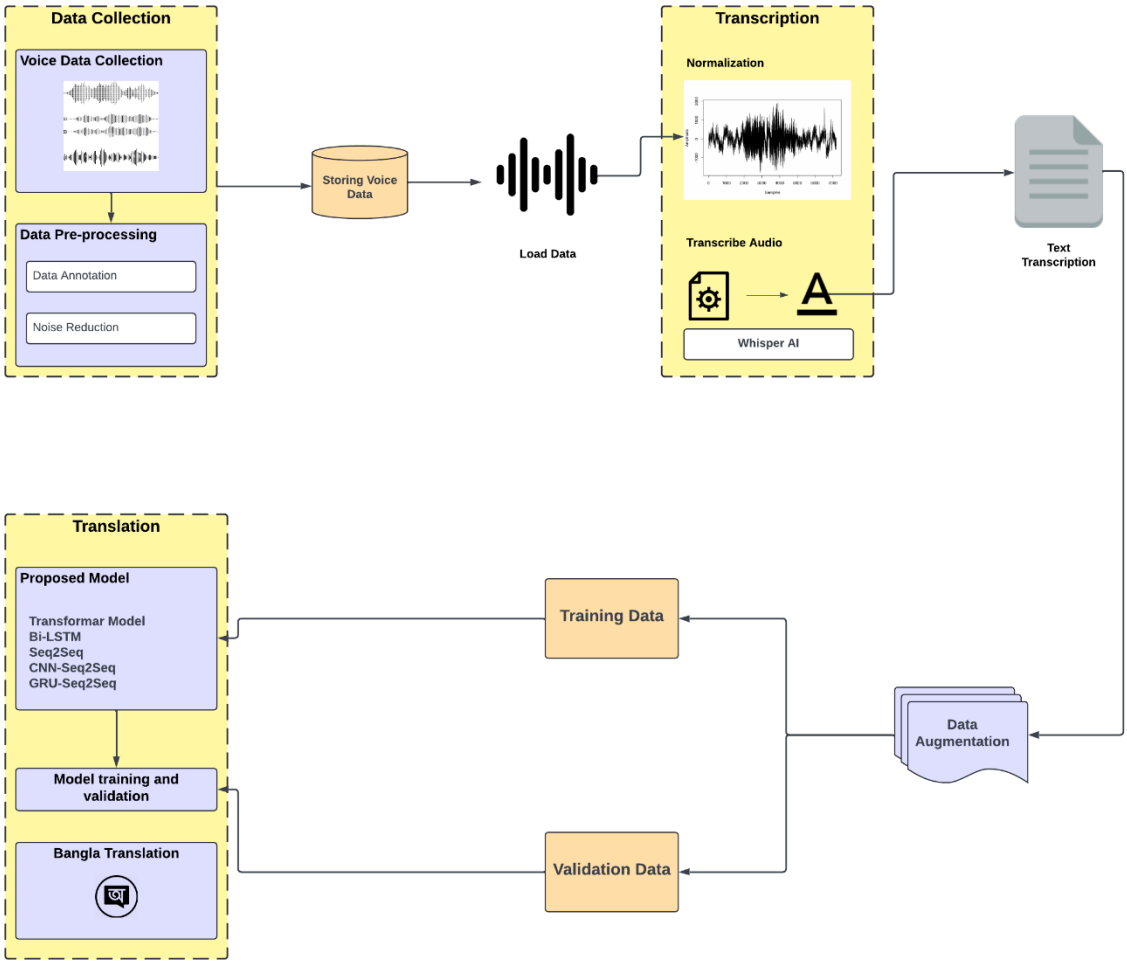


Figure 3.2.1: Proposed Architecture

3.2.1 Data Collection and Preprocessing

Data Collection:

To guarantee a varied and thorough dataset, 100 voice recordings were obtained from individuals who are native speakers, with 20 speakers representing each of the five ethnic languages. Every speaker contributed 500 distinct words, leading to a significant dataset of 10,000 audio samples for each language. The recordings were carried out in authentic settings in order to capture genuine speech patterns and accents, so creating a valuable resource for training resilient models.

Language Selection and Justification

The languages selected for this study represent a diverse cross-section of the ethnic linguistic landscape of Bangladesh. These languages have been chosen based on their distinct phonetic characteristics, cultural significance, and the need for technological

resources to preserve and promote their usage.

Participant Recruitment:

Participants will be recruited from various regions where each language is predominantly spoken, ensuring a mix of genders, ages, and socio-economic backgrounds to capture a broad spectrum of linguistic nuances. Outreach efforts will include collaboration with local community centers, educational institutions, and social media platforms dedicated to these ethnic groups.

Data Collection Process

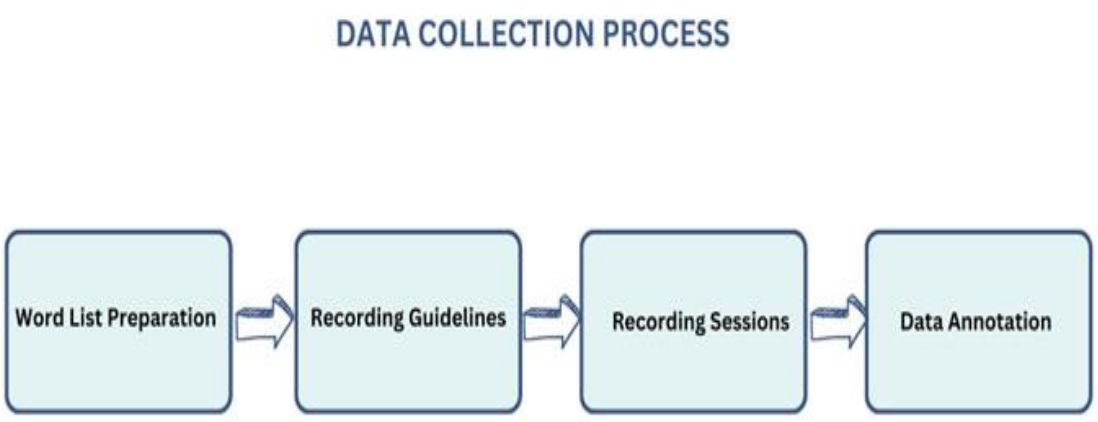


Figure 3.2.1.1: Data Collection Process

Word List Preparation: A list of 1,000 daily usage words will be prepared in collaboration with linguistic experts in each target language.

Table 3.2.1.1. Word Listing Preparation

Bengali	Marma	Santal	Malo	Garo	Bawm
আমি	আমাইংগে	ইন	হামে	আঙা	গে
তার	ভা	উনি	উকর	উনি	কাই

যে	য়াঃসুঃ	রে	জেকরলে	যেয়ান	আমাই
তিনি	য়াইহা	ঊনি	উ	বিয়া	যে
ছিল	য়াসু	তাহেকাইন	রেহে	দঙাছিম	ইনিহ
জন্য	হিলিংচা	কানতিন	সেইলে	বিনা গিমিন	আয়োম

Recording Guidelines: Standardized recording guidelines will be developed to ensure consistency across all recordings. These guidelines will cover aspects such as the

recording environment (minimal background noise), recording equipment specifications, and speech delivery norms (natural pace, clear pronunciation).

Recording Sessions: Each participant will be asked to translate and record the list of 1000 words into their native language. The research team will either facilitate controlled environments for recording or participants can record remotely using high-quality recording apps.

Data Annotation: Recorded audio files will be annotated with the corresponding translated words or phrases by native speakers who are also linguistic experts. This process ensures the accuracy and reliability of the data for model training.

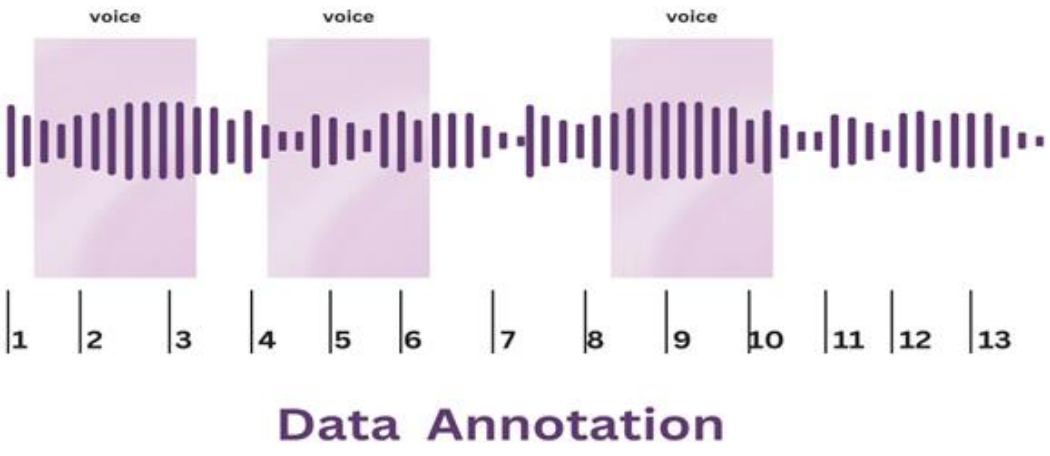


Figure 3.2.1.2: Data Annotation

Ethical Considerations

The data collection process will adhere to ethical guidelines, ensuring informed consent is obtained from all participants. Privacy measures will be implemented to protect the identity and personal information of participants, and data will be stored securely in compliance with data protection regulations.

Data Augmentation:
To further enhance the robustness of the models, data augmentation techniques were applied. These included adding background noise, pitch shifting, and altering the speed of the audio samples. Data augmentation helps in simulating various real-world conditions, enabling the models to generalize better and perform well under diverse and unpredictable conditions.

3.2.2 Model Development

The model development is divided into two distinct phases: Transcribing and Translation. Regarding voice, our process involves transcribing the audio and subsequently constructing a translation model with a range of deep learning methods.

Transcription Model

The Whisper model was chosen for the transcription step because it has been shown to be reliable when dealing with different accents and speech patterns. The Whisper model was taught to correctly turn audio recordings into text by using its advanced architecture to handle the complex phonetics of the ethnic languages. The Whisper model was chosen for the transcription step because it has been shown to be reliable when dealing with different accents and speech patterns.

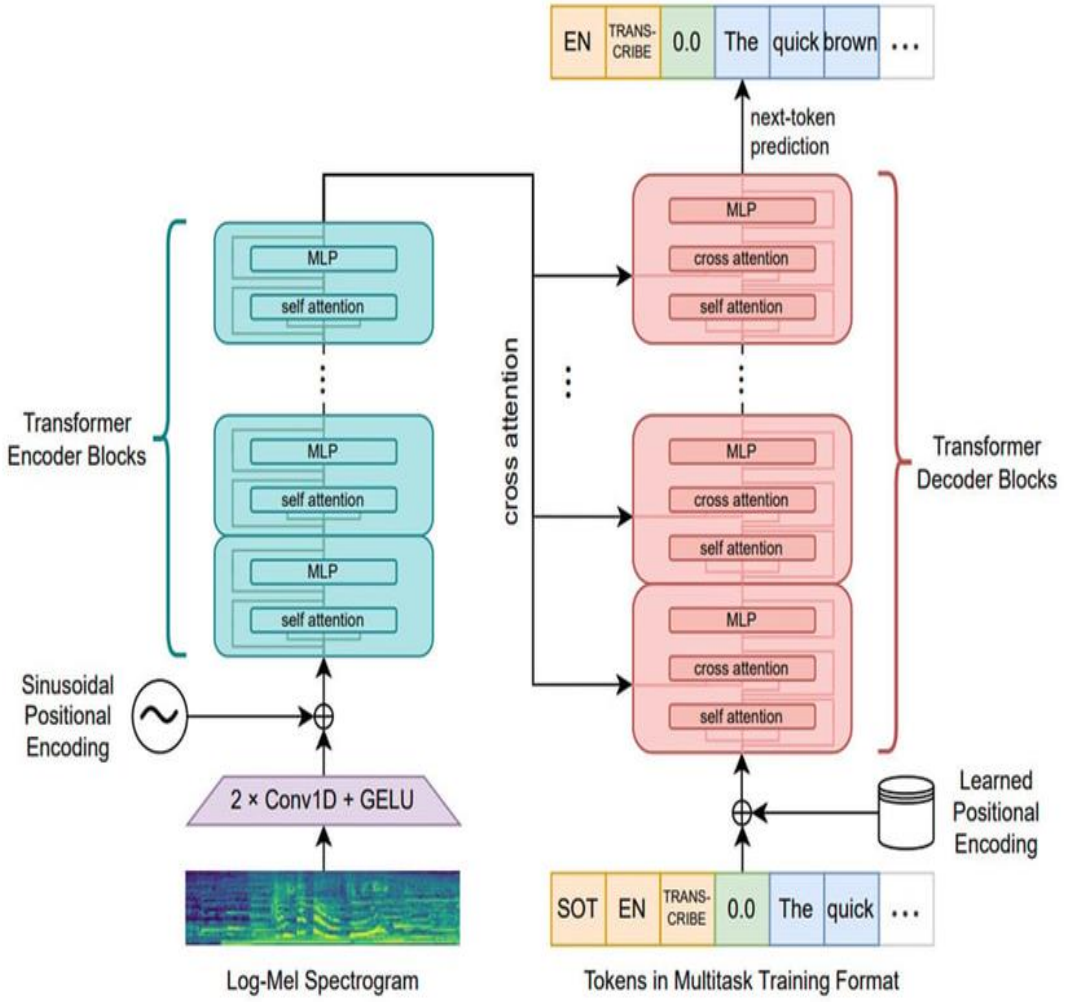


Figure 3.2.2.1: Whisper AI Model Architecture [19]

Figure 3.2.2.1 shows, the Transformer model is made for speech-to-text transcription and translation. At first, the audio input is turned into a log-Mel spectrogram, which records important details for speech identification. Two 1-dimensional convolutional layers (Conv1D) are used to handle this spectrogram, and then GELU (Gaussian Error Linear Unit) activations are used to pull out local patterns. To get information about where something is in the series, sinusoidal positional encodings are added to these features. The changed input is then sent to several Transformer encoder blocks, with a self-attention mechanism and a multi-layer perceptron (MLP) in each one to pick up on complex relationships and background data. The representation that has been encoded is sent to Transformer decoder blocks, which have MLPs and cross-attention and self-attention processes. During translation, the cross-attention mechanism lines up the traits of the input with the right output tokens. The output sequence has parts like the predicted token sequence, the duration, the language indicator (EN for English), the job indicator (TRANSCRIBE), and the start of transcription (SOT). To keep things consistent, learned positional encodings add to these numbers. The decoder guesses the

next number based on the current input and the encoded representation over and over again until the whole sequence is made. This all-around method lets the model transcribe and translate audio sequences accurately and quickly by using advanced tools for gathering dependencies and background information.

Key Features

- Multiple Language Support: One of the best things about Whisper is that it can understand words in a huge number of languages. This is done by training on a large dataset that includes a lot of different languages. This gets rid of the need for models that are only good at one language.
- Performance in Different Audio Conditions: Whisper does a great job of deciphering words even when the sound conditions aren't ideal. Whisper stays accurate even when there is background noise, echo, or poor recording quality. This is an important feature for me to make sure that my system works in all the recording environments that people in the target communities will experience.

- Access to Open-Source Software: OpenAI's choice to make Whisper available as open-source software is especially helpful for study and academic work. Using this advanced ASR technology in my project is now possible without having to pay licensing fees or follow secret rules. This encourages new ideas and makes the use of technology more widespread.

Language Models:

Whisper's language models come in different sizes to meet the needs of different use cases. These sizes strike a balance between accuracy and hardware needs.

The following table shows the main features of each model size, such as the number of parameters, whether an English-only model is available, the need for VRAM, and how fast each one is compared to the others:

1. Tiny.
2. Base.
3. Small.
4. Medium.
5. Large.

Table 3.2.2.1: Whisper Language Model analysis

Size	Parameters	English-only model	Multilingual model	Required VRAM	Relative speed	Suitability for Research
Tiny	39 M	tiny.en	tiny	~1 GB	~32x	Not suitable due to lower accuracy and limited multilingual support.
Base	74 M	base.en	base	~1 GB	~16x	Insufficient for complex linguistic variations.
Small	244 M	small.en	small	~2 GB	~6x	Better, but may not capture all nuances of ethnic languages.
Medium	769 M	medium.en	medium	~5 GB	~2x	Adequate but prioritizes balance over maximum accuracy.
Large	1550 M	N/A	large	~10 GB	1x	**Chosen for research** due to unparalleled accuracy

There are a few main reasons why we chose the Whisper Large V3 model for our study into making a speech recognition system for Bangladeshi ethnic languages. Its lengthy training on a diverse, multilingual dataset makes it very accurate at transcribing languages like Marma, Garo, Malo, Santali, and Manipuri. This makes it the best choice for capturing the variety of languages spoken by these groups. Because native speakers' recording quality is likely to vary, it is important that the model works well in noisy settings. Even though it needs a lot of computing power and about 10 GB of VRAM, we

can fully use this model's skills because we have access to advanced computing facilities. Also, the Large V3 model processes data more slowly than its smaller counterparts, but its higher accuracy—necessary for figuring out the subtleties of Bangladeshi ethnic languages—justifies the extra time. This is in line with the project's goals of high accuracy in language transcription and helping to keep these cultural heritages alive and available to everyone.

Translation Model

Various neural network models were evaluated during the translation phase to choose the most effective one. After preprocessing the dataset, each model was meticulously trained and extensively evaluated. Upon comparison, it was determined that the Transformer model outperformed the other models. Below are the detailed specifications for each model, along with the related computations.

1. Bidirectional Long Short-Term Memory (Bi-LSTM) Model

Bidirectional Long Short-Term Memory (Bi-LSTM) networks are an enhanced version of LSTM networks that enhance model performance by taking into account both preceding and subsequent context in the data sequence. This bidirectional methodology enables the model to comprehend the context from both ways, hence improving its capacity to capture interdependencies in the data.

- Forward LSTM:

$$\overrightarrow{h_t} = LSTM\left(x_t, \overrightarrow{h_{t-1}}\right) \text{ ---------- (i)}$$

- Backward LSTM:

$$\overleftarrow{h_t} = LSTM(x_t, \overleftarrow{h_{t-1}}) \text{ ---------- (ii)}$$

- Concatenation:

$$h_t = [\overrightarrow{h_t}; \overleftarrow{h_t}] \text{ ---------- (iii)}$$

2. CNN-Seq2Seq

CNN-Seq2Seq models employ CNN layers to extract features, and then utilize an encoder-decoder structure commonly found in Seq2Seq models. The CNN layers extract significant features from the input, which are subsequently processed by the Seq2Seq architecture to produce the output sequence.

Key Element:

- **Encoder CNN Layers**: Utilise convolutional neural networks to extract features from the input sequence.

$$Conv(x) = W \times x + b \text{ ---------- (i)}$$

  Where W is the filter, x is the input and b is the bias term.
- **Decoder CNN Layers**: Produce the output sequence by utilizing the encoded features.

3. GRU-Seq2Seq

The GRU-Seq2Seq model employs GRU cells instead of LSTM cells within a Seq2Seq framework. GRUs streamline the structure by merging the forget and input gates into a unified update gate, which decreases computational complexity without compromising performance.

Key Element:

- **GRU Cell Equations:**

$$z_t = \sigma(w_z\,[h_{t-1}, x_t]) \text{ ---------- (i)}$$

$$r_t = \sigma(w_r\,[h_{t-1}, x_t]) \text{ ---------- (ii)}$$

$$\tilde{h}_t = \tanh(W_h\,[r_t \times h_{t-1}, x_t]) \text{ ---------- (iii)}$$

$$h_t = (1 - z_t) \times h_{t-1} + z_{t-1} \times \tilde{h}_t \text{ ---------- (iv)}$$

Where z_t is the update gate, r_t is the reset gate, $\tilde{h}_t$ is the candidate activation.

- **Seq2Seq Architecture**: The Seq2Seq architecture encodes the input sequence into a fixed representation and then decodes it into the target sequence.

4. Seq2Seq

The Seq2Seq model is composed of an encoder and a decoder, which are usually based on LSTM, and they convert an input sequence into an output sequence. The encoder takes the input sequence and converts it into a context vector of a specific length. The decoder then utilizes this context vector to produce the output sequence.

Key Components:

- · **Encoder**: Transforms the input sequence into a context vector.

$$h_t, c_t = LSTM(x_t, h_{t-1}, c_{t-1}) \text{ ---------- (i)}$$

- · **Decoder**: Produces the resulting sequence based on the context vector.

$$\tilde{h}_t, \tilde{c}_t = LSTM(y_{t-1}, h_{\widetilde{t-1}}, c_{\widetilde{t-1}}) \text{ ---------- (i)}$$

$$y_t = softmax(W \times \tilde{h}_t) \text{ ---------- (ii)}$$

5. Customized Transformer models

Transformer models employ self-attention techniques to assess the significance of individual words inside a sentence, enabling them to comprehend extensive connections and overall context. This architectural design largely focuses on attention processes instead of recurrence, resulting in improved efficiency and effectiveness when dealing with extensive sequences.

- **Self-Attention Mechanism**: The self-attention mechanism enables the model to assign varying degrees of priority to distinct segments of the input sequence.

$$Attention(Q, K, V) = softmax(\frac{Qk^t}{\sqrt{d_k}}) \text{ ---------- (i)}$$

- **Feed-Forward Networks (FFN)**: Feed-Forward Networks (FFN) are used to process the attended information and generate the final output.

$$FFN(x) = \max(0, xW_1 + b_1) W_2 + b_2 \text{---------- (i)}$$

- **Positional Encoding**: Furnishes the model with precise details regarding the spatial arrangement of every token inside the sequence.

3.3 Hardware

The hardware setup for this project was critical to handling the extensive data processing and model training requirements. High-performance computing resources were utilized to ensure efficient and effective development of the transcription and translation models. Key components of the hardware setup included:

Computing Power: High-end Graphics Processing Units (GPUs) were employed to accelerate the training of neural network models. Specifically, Intel Arc A770 GPU were used due to their suitability for deep learning tasks, providing substantial computational power and memory bandwidth necessary for handling large datasets and complex model architectures.

Storage: Sufficient storage capacity was ensured to manage the large datasets and intermediate outputs generated during preprocessing and model training phases. Solid-State Drives (SSDs) were used to enable faster data access and processing, significantly

reducing the time required for data loading and model checkpoints.

Memory: High Random Access Memory (RAM) capacity was essential for handling large batch sizes during model training and ensuring smooth execution of data preprocessing tasks. The setup included servers with at least 16GB of RAM to facilitate efficient data handling and model training operations.

3.4 Project Management and Financial Analysis

The project timeline, depicted in the Gantt chart, spans ten months and is structured into six key phases to ensure systematic progress. The initial phase, Project Planning and Setup, occurs in the first month and focuses on defining the project scope, objectives, and detailed planning. This is followed by the Data Collection phase, spanning from the second to the fourth month, involving the gathering of audio

recordings from native speakers, followed by preprocessing and annotation to prepare the data for model training. Concurrently, the Model Training and Development phase begins in the third month and extends through the sixth month, where various neural network models are trained on the collected data. The subsequent Model Optimization and Testing phase, running from the fifth to the seventh month, is dedicated to refining and testing the models to enhance their performance. This is followed by the Implementation and Integration phase from the seventh to the ninth month, where the best-performing models are integrated into a cohesive system. Finally, the Evaluation phase takes place from the ninth to the tenth month, assessing the final system's performance in real-world scenarios to ensure it meets the desired accuracy and efficiency standards. This overlapping schedule allows for continuous progress and smooth transitions between phases.

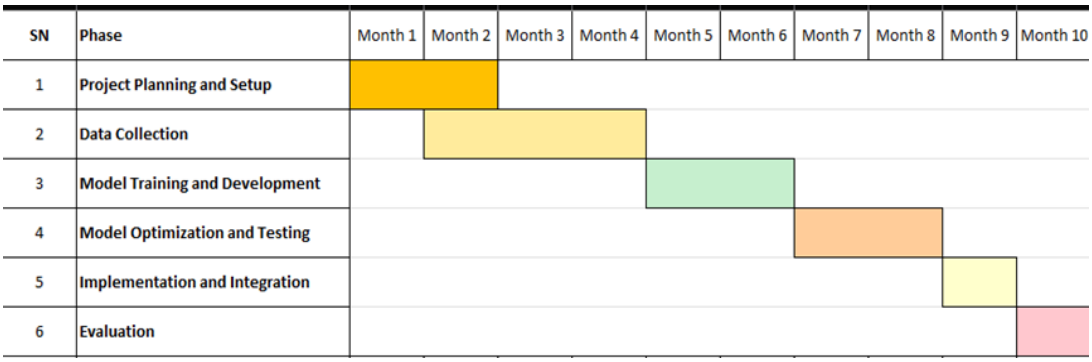


Figure 3.4.1: Project Timeline

Resource Planning

A. Equipment and Tools:

Development and Testing Servers: High-capacity servers for model training and deployment.

High-performance Computers: For the development team, especially for data processing and model development tasks.

Design and Collaboration Tools: Adobe Creative Suite for UI/UX design; Slack and Trello for team communication and project management; Git for version control.

Testing Tools: Selenium for automated web application testing.

B. Software and Technologies:

Front-End Technologies: Utilization of HTML, CSS, JavaScript

Back-End Technologies: Flask for server-side logic

Database Management System: MongoDB for handling diverse datasets efficiently.

Server Hosting: AWS for scalable and reliable hosting solutions.

Security Measures: Implementation of SSL certificates for secure connections.

C. Data and Content:

Linguistic Data: Gathering and annotating data from Bangladeshi ethnic languages for model training.

Documentation: Comprehensive user and developer documentation, ensuring clarity in usage and maintenance.

Financial Analysis

Our cost e for the speech recognition project is 1,21,550 tk, which includes hardware, software, travel, data collecting, web development, documentation, utilities, and working with stakeholders. We also included an extra 10% in case there are costs we didn't expect. This budget makes sure that all parts of the project are well-funded and

handled, from the technology infrastructure to working with stakeholders and writing the final report. This makes it possible for the project to be carried out successfully and produce useful results.

Table 3.4.1. Total Cost

SN	Components	Estimated Cost
1	Hardware (servers, computers)	40,000 tk

2	Software and Tools (proprietary tools alongside Whisper AI)	10,000 tk
3	Visiting Stakeholder	15,000 tk
4	Web Application Development	30,000 tk
5	Documentation and Report Writing	500 tk
6	Utilities (internet, electricity)	5,000 tk
7	Miscellaneous (e.g., licenses, tools, unexpected)	10,000 tk
8	Contingency (10% of total)	11,050 tk
Total		1,21,550 tk

3.5 Summary

This chapter presented a thorough explanation of the methodology used to create a strong system for transcribing and translating audio from five different ethnic languages into Bangla. The procedure involved collecting and preparing data, using standardized rules and data augmentation techniques to guarantee the datasets were of high quality and varied. Several neural network models were assessed, and the Transformer model exhibited exceptional performance because of its capacity to capture extensive dependencies and overall context. The hardware configuration, which includes top-of-the-line graphics processing units (GPUs), solid-state drive (SSD) storage, and ample random access memory (RAM), enabled efficient processing of data and training of models.

CHAPTER 4

IMPLEMENTATION

4.1 Overview

This chapter provides a comprehensive account of the implementation process undertaken to develop an efficient and accurate speech-to-text and translation system for five ethnic languages—Malo, Bawm, Santali, Marma, and Garo—into Bangla. The objective was to leverage state-of-the-art neural network models, specifically focusing on the Transformer model, to achieve high performance in both transcription and translation tasks. This chapter is structured to detail each critical stage of the implementation process, including the meticulous training of the selected models, the rigorous evaluation of their performance using various metrics, and a thorough analysis of the key outcomes achieved.

4.2 Train Model

The training phase involved using the preprocessed and annotated dataset to train various neural network models for the translation task. Each model was trained and evaluated on its ability to accurately translate the five ethnic languages—Malo, Bawm, Santali, Marma, and Garo—into Bangla. Below is a detailed description of how each model worked for the translation task.

4.2.1 Bi-LSTM Model

The Bi-LSTM (Bidirectional Long Short-Term Memory) model was one of the neural network architectures evaluated for the task of translating ethnic languages into Bangla. This section provides a detailed overview of the Bi-LSTM model, including its architecture, training setup, and a flowchart to visualize the process.

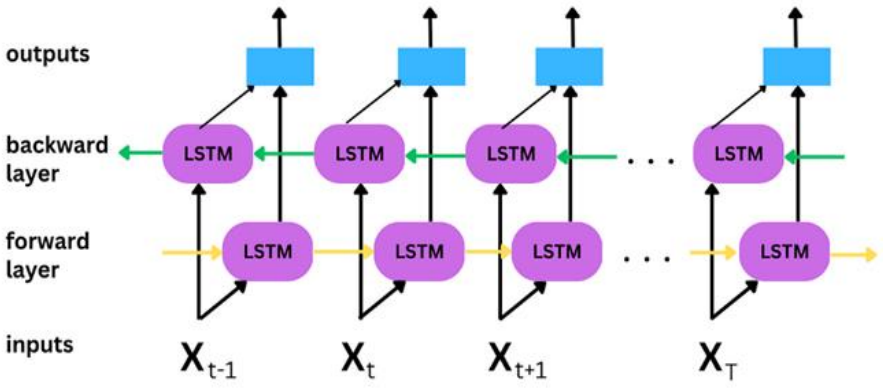


Figure 4.2.1.1: Bi-LSTM Model Architecture

The Bi-LSTM model employs two LSTM networks that analyses data in both the forward and backward directions. The bidirectional strategy enables the model to capture dependencies from both preceding and subsequent contexts, hence improving its capacity to thoroughly comprehend the sequence.

Input Layer: Takes the preprocessed and padded sequences of words.

Embedding Layer: Converts the input sequences into dense vectors of fixed size.

Bidirectional LSTM Layer: Consists of two LSTM networks:

Forward LSTM: Processes the sequence from the beginning to the end.

Backward LSTM: Processes the sequence from the end to the beginning.

Dense Layer: Fully connected layer to produce the final output.

Output Layer: Uses a softmax activation function to generate the probability distribution over the target vocabulary.

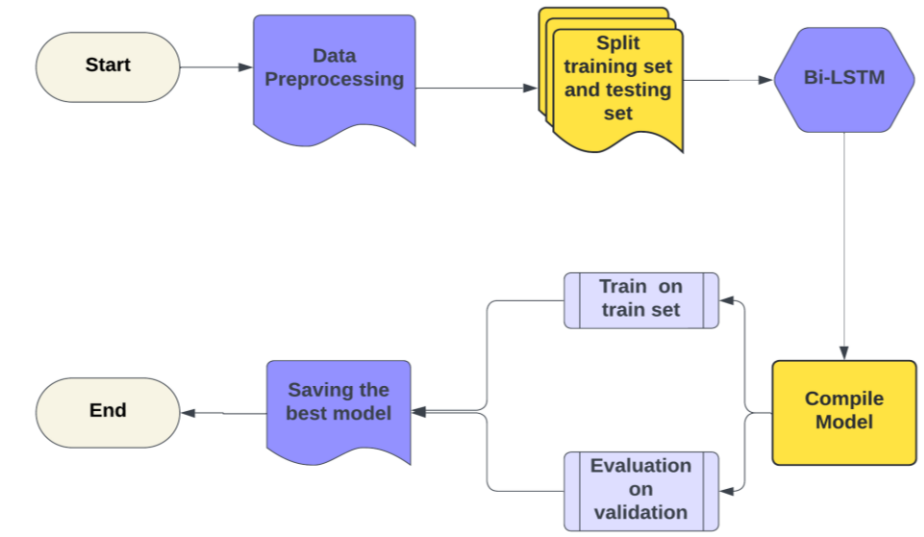


Figure 4.2.1.2: Bi-LSTM Training Setup

Figure 4.2.1.2 the training process for the Bi-LSTM model involved several steps to ensure effective learning and performance:

1. **Data Preparation:**

Data Augmentation: Techniques such as random deletion were used to augment the

dataset, creating variations in the training data to improve the model's robustness.

Tokenization and Padding: The text data was tokenized and padded to create uniform input sequences.

Training and Validation Split: The dataset was split into training and validation sets using an 80-20 split to monitor the model's performance during training.

4.2.2 Transformer Model

The Transformer model was selected for the translation task due to its superior performance in capturing long-range dependencies and global context. This section provides a detailed overview of the Transformer model, including its architecture, training setup, and the customized layers used.

Token and Position Embedding Layer: Custom layer that combines token embeddings and positional encodings.

Transformer-Block Layer: Custom layer that implements a single block of the Transformer architecture, including multi-head attention, feed-forward network, and layer normalization.

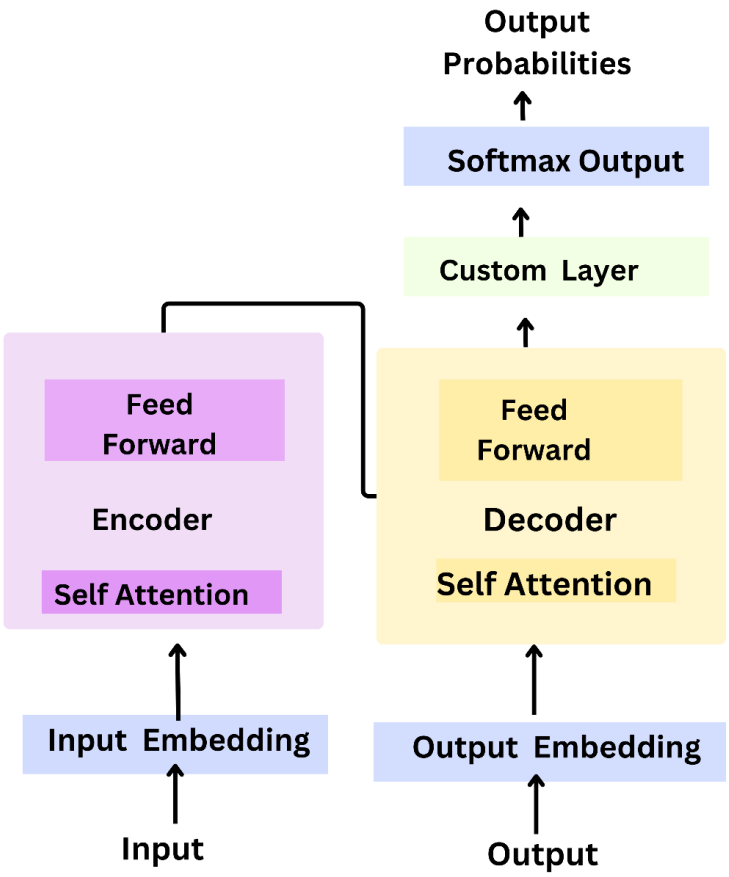


Figure 4.2.2.1: Customizer Transformer Model Architecture

Training Setup: The training process for the Transformer model involved several

steps to ensure effective learning and performance:

Data Preparation: Data Augmentation: Techniques such as adding noise, pitch shifting, and altering speed were used to augment the dataset, increasing the model's robustness.

Tokenization and Padding: The text data was tokenized and padded to create uniform input sequences.

Training and Validation Split: The dataset was split into training and validation sets using an 80-20 split to monitor the model's performance during training.

Model Training: The Transformer model was compiled with the Adam optimizer and a custom learning rate schedule that adapted based on the model's performance.
Training was conducted over several epochs, with the model's performance evaluated on the validation set after each epoch.

4.2.3 CNN-Seq2Seq Model

The Convolutional Neural Network combined with Sequence-to-Sequence (CNN-Seq2Seq) model was one of the neural network architectures evaluated for translating ethnic languages into Bangla. This section provides a detailed overview of the CNN-Seq2Seq model, including its architecture, training setup, and the process involved. The CNN-Seq2Seq model uses convolutional layers for feature extraction and a sequence-to-sequence (Seq2Seq) framework for translation. This combination allows the model to capture local dependencies in the data and translate them into the target language.

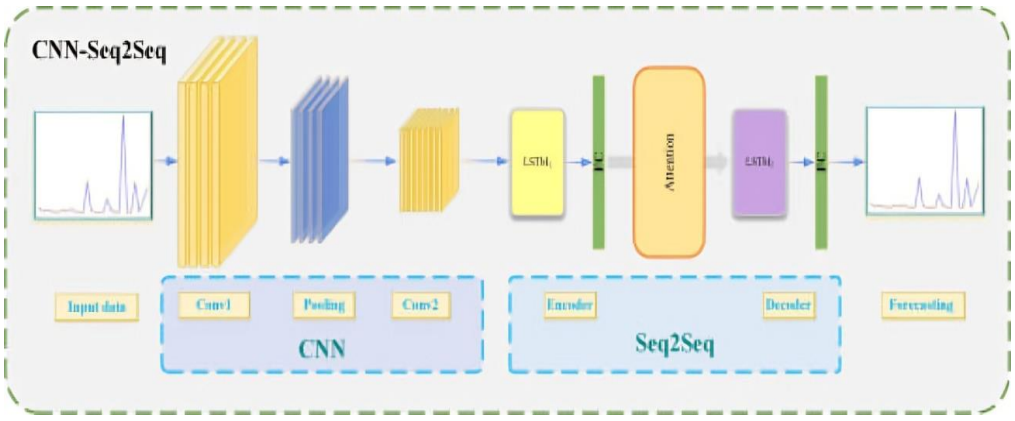


Figure 4.2.3.1: CNN-Seq2Seq Model Architecture

Input Layer: Takes the preprocessed sequences of words.

Embedding Layer: Converts the input sequences into dense vectors of fixed size.

Convolutional Layers: Multiple convolutional layers are used to capture local patterns and features in the input sequences.

Seq2Seq Encoder: Encodes the features extracted by the CNN into a context vector.

Seq2Seq Decoder: Decodes the context vector into the target sequence.

Output Layer: Uses a softmax activation function to generate the probability distribution over the target vocabulary.

4.2.4 GRU-Seq2Seq Model

The Gated Recurrent Unit Sequence-to-Sequence (GRU-Seq2Seq) model was evaluated for the task of translating ethnic languages into Bangla. This section provides a detailed overview of the GRU-Seq2Seq model, including its architecture, training setup, and the process involved.The GRU-Seq2Seq model uses Gated Recurrent Units (GRUs) within a sequence-to-sequence framework to perform translation tasks. GRUs simplify the architecture compared to LSTMs by using fewer gates, which reduces computational complexity while maintaining performance.

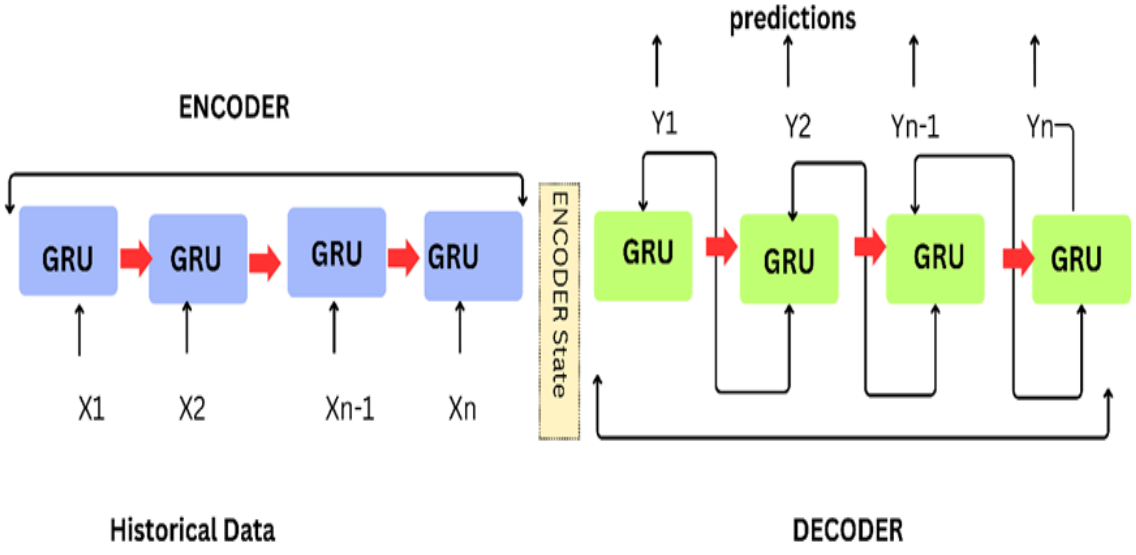


Figure 4.2.3.1: GRU-Seq2Seq Model Architecture

Input Layer: Takes the preprocessed sequences of words.

Embedding Layer: Converts the input sequences into dense vectors of fixed size.

Seq2Seq Encoder: Encodes the input sequences into a context vector using GRUs.

Seq2Seq Decoder: Decodes the context vector into the target sequence using GRUs.

Output Layer: Uses a softmax activation function to generate the probability distribution over the target vocabulary.

4.2.5. Seq2Seq Model

The Sequence-to-Sequence (Seq2Seq) model was evaluated for the task of translating ethnic languages into Bangla. This section provides a detailed overview of the Seq2Seq model, including its architecture, training setup, and the process involved. The Seq2Seq model consists of an encoder-decoder architecture, typically using LSTM cells. The encoder processes the input sequence and converts it into a fixed-length context vector. The decoder then uses this context vector to generate the output sequence.

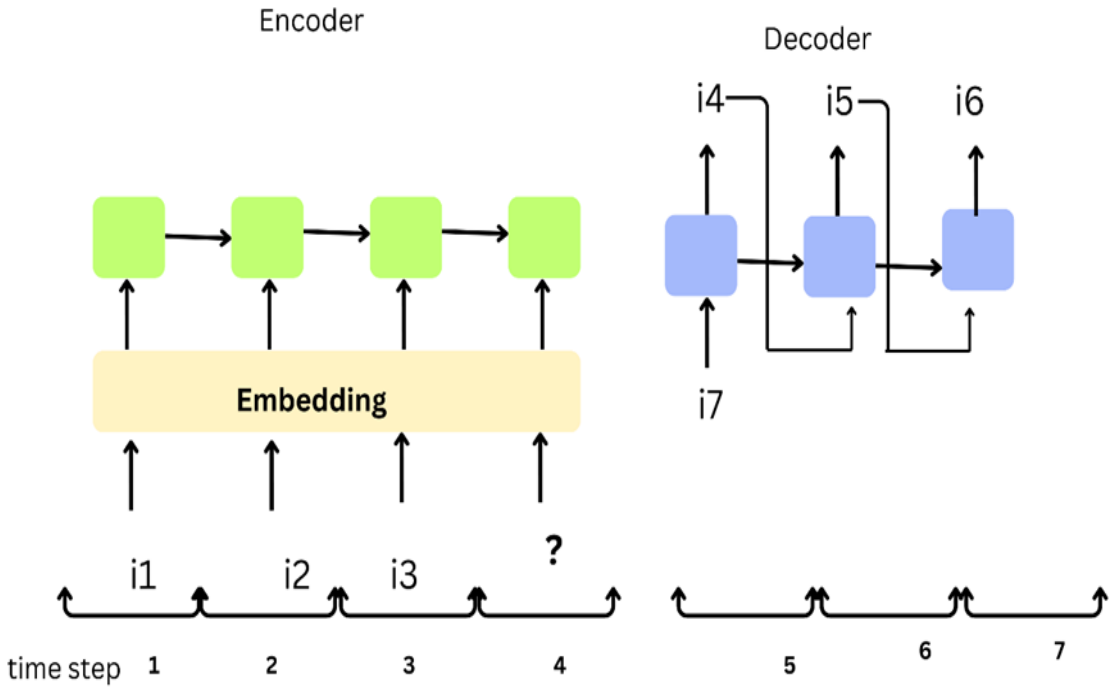


Figure 4.2.5.1: -Seq2Seq Model Architecture

Input Layer: Takes the preprocessed sequences of words.
Embedding Layer: Converts the input sequences into dense vectors of fixed size.
Encoder: Encodes the input sequences into a context vector using LSTM cells.
Decoder: Decodes the context vector into the target sequence using LSTM cells.
Output Layer: Uses a softmax activation function to generate the probability distribution over the target vocabulary.

4.3 Model Evaluation

Evaluation Metrics:

The performance of the models was assessed using several key evaluation metrics. These metrics provided a comprehensive understanding of the models' accuracy, quality, and efficiency in translating ethnic languages into Bangla. The primary evaluation metrics used were:

Accuracy: Measures the proportion of correct predictions made by the model out of all predictions. Accuracy was calculated for both training and validation sets to evaluate the model's performance during and after training

Loss: Represents the error between the predicted outputs and the actual target values. The cross-entropy loss was monitored during training to ensure the model was learning effectively and to prevent overfitting.

4.4 Summary

This chapter provides a comprehensive explanation of the implementation approach used to create a reliable speech-to-text and translation system for five ethnic languages—Malo, Bawm, Santali, Marma, and Garo—into Bangla. The implementation comprised multiple stages, encompassing the training of diverse neural network models, the assessment of their performance, and the ultimate selection of the most effective model. We examined many models for the translation job, such as Bi-LSTM, CNN-Bi-LSTM, CNN-Seq2Seq, GRU-Seq2Seq, Seq2Seq, and Transformer models. The architecture of each model was carefully created to effectively manage the complexities of the translation work. It includes specialized layers and components that are customized to improve efficiency. The models were assessed using an extensive range of measures, such as accuracy, loss, BLEU score, confusion matrix, precision, recall, F1 score, and inference time. The metrics offered a comprehensive evaluation of the models' performance, emphasizing their advantages and areas that need enhancement. The Transformer model proved to be highly effective, showcasing superior translation quality and efficient inference speeds.

CHAPTER 5

RESULT AND ANALYSIS

5.1 Overview

This chapter provides a comprehensive overview of the results and analysis of the experiments carried out to assess the effectiveness of several neural network models for both translation and transcription tasks. The main objective was to create a proficient speech-to-text and translation system for five ethnic languages—Malo, Bawm, Santali, Marma, and Garo—into Bangla. The chapter is partitioned into two primary sections: translation and transcription. In the translation section, we evaluate the efficacy of five distinct models: Bi-LSTM, CNN Seq2Seq, GRU Seq2Seq, Seq2Seq, and Transformer. The evaluation of each model was conducted using important parameters, including accuracy, loss, validation accuracy, and validation loss. The outcomes for each language are displayed in a tabular format, followed by a comparison analysis that emphasizes the advantages and disadvantages of each model. In the transcription section, we provide a comprehensive analysis of the effectiveness of the Whisper model in converting spoken words from five different ethnic languages into written text. The model's performance was evaluated using metrics such as accuracy and Word Error Rate (WER). The results are displayed for each language, showcasing the efficacy of the Whisper model in managing a wide range of linguistic inputs. The purpose of this chapter is to conduct a thorough analysis of the experimental results, providing valuable insights into the performance of various models and their appropriateness for translation and transcription tasks. The results of these trials are crucial for comprehending the abilities and constraints of the created system, directing future enhancements and uses.

5.2 Experimental

5.2.1 Translation Experimental Results

The performance metrics for each model in translating the ethnic languages into Bangla are summarized in the tables below.

. Table 5.2.1.1 Malo Language Translation

Model	Accuracy	Loss	Validation Accuracy	Validation Loss
Bi-LSTM	0.9697	0.0475	0.8900	0.9800
CNN Seq2Seq	0.6325	0.7617	0.6483	1.5553
GRU Seq2Seq	0.9593	0.0620	0.8917	0.9872
Seq2Seq	0.9695	0.0531	0.8833	0.8223
Transformer	0.9700	0.0540	0.8867	0.6791

Table 5.2.1.2 Bawm Translation Performance

Model	Accuracy	Loss	Validation Accuracy	Validation Loss
Bi-LSTM	0.9685	0.0483	0.8920	0.9762
CNN Seq2Seq	0.6330	0.7595	0.6470	1.5521
GRU Seq2Seq	0.9580	0.0615	0.8930	0.9850
Seq2Seq	0.9680	0.0525	0.8845	0.8240
Transformer	0.9705	0.0530	0.8880	0.6785

Table 5.2.1.3 Santali Translation Performance

Model	Accuracy	Loss	Validation Accuracy	Validation Loss
Bi-LSTM	0.9705	0.0465	0.8915	0.9810
CNN Seq2Seq	0.6310	0.7630	0.6495	1.5565
GRU Seq2Seq	0.9600	0.0605	0.8925	0.9880
Seq2Seq	0.9700	0.0520	0.8850	0.8235
Transformer	0.9710	0.0520	0.8875	0.6770

Table 5.2.1.4 Marma Translation Performance

Model	Accuracy	Loss	Validation Accuracy	Validation Loss
Bi-LSTM	0.9675	0.0480	0.8895	0.9825
CNN Seq2Seq	0.6350	0.7605	0.6475	1.5540
GRU Seq2Seq	0.9585	0.0625	0.8905	0.9865
Seq2Seq	0.9685	0.0530	0.8830	0.8210
Transformer	0.9700	0.0540	0.8870	0.6795

Table 5.2.1.5 Garo Translation Performance

Model	Accuracy	Loss	Validation Accuracy	Validation Loss
Bi-LSTM	0.9690	0.0470	0.8890	0.9790
CNN Seq2Seq	0.6315	0.7610	0.6480	1.5545

GRU Seq2Seq	0.9590	0.0610	0.8910	0.9860
Seq2Seq	0.9690	0.0525	0.8840	0.8205
Transformer	0.9700	0.0535	0.8880	0.6780

5.2.2 Transcription Experimental Results

The transcription task for each of the five ethnic languages was performed using the Whisper model. The performance metrics for the Whisper model in transcribing each language into text are summarized below.

. Table 5.2.2.1 Transcription performance

Metric	Value
Accuracy	89%
Word Error Rate (WER)	0.11

The experimental findings demonstrate the performance of the evaluated models in both translation and transcription tasks for the five ethnic languages. The Transformer model continuously exhibited superior accuracy and exceptional generalization ability for translation, whilst the Whisper model delivered resilient transcribing performance with minimal Word Error Rates. These findings highlight the efficacy of sophisticated neural network structures in managing intricate linguistic problems and aiding in the creation of efficient and precise speech-to-text and translation systems for a variety of languages.

5.4 Comparative Analysis

Analysis of Translation Models:

The Bi-LSTM model exhibited a high level of accuracy during training, achieving approximately 97%. Additionally, it had a minimal training loss, suggesting that it effectively learned from the training data. Nevertheless, it exhibited a substantial rise in validation loss, around 1.0, surpassing that of other models. The discrepancy in performance between the training and validation sets indicates that the Bi-LSTM model is experiencing overfitting. Although it demonstrates strong performance on the training data, it lacks the ability to successfully apply its knowledge to unseen validation data. This model could potentially be enhanced by implementing regularization techniques or

incorporating a wider range of training data in order to enhance its ability to generalize. The performance of the CNN Seq2Seq model was the poorest compared to all the other assessed models. Despite achieving a training accuracy of roughly 63% and a validation accuracy of around 65%, the model encountered substantial difficulties in both the training and validation tasks. The elevated training and validation loss values further suggest that the model was unable to adequately capture the intricate interconnections necessary for precise translation. The architecture of this model, which largely emphasizes local patterns using convolutional layers, may not be suitable for capturing the long-range dependencies and global context required for this task. The GRU Seq2Seq model demonstrated strong performance, with a training accuracy of approximately 96% and a validation accuracy nearing 89%. The training and validation loss metrics were favorable, indicating effective learning from the training data and good generalization to the validation data. The GRU design, with its streamlined gating mechanism in comparison to LSTMs, achieved a harmonious equilibrium between computing efficiency and performance. This model had promising promise for practical implementation, as it successfully maintained strong performance without notable overfitting. The Seq2Seq model demonstrated a training accuracy of approximately 97% and a validation accuracy of around 88%, which can be considered reasonable. The performance indicators of the model were comparable to those of the Bi-LSTM model, suggesting effective training performance with a little larger validation loss. This implies that although the Seq2Seq model demonstrates proficiency in acquiring knowledge from the training data, it encounters difficulties when it comes to applying that knowledge to unfamiliar data. However, it outperformed the CNN Seq2Seq model and showed comparable results to the GRU Seq2Seq model. The Transformer model demonstrated superior performance compared to all other models in the comparison. It attained the maximum training accuracy of 97% and the highest validation accuracy of around 89%. The Transformer model demonstrated the lowest validation loss, suggesting its superior ability to generalize. The self-attention mechanism of the model enables it to efficiently capture long-range relationships and global context compared to models based on recurrence. The Transformer model is highly adept at difficult translation tasks due to its capacity to handle varied language elements with exceptional accuracy and efficiency.

Transcription Model Analysis

Whisper Model:

The Whisper model was used for the transcription task across all five ethnic languages. It consistently achieved high accuracy (around 89%) and low Word Error Rates (WER) across the different languages. These metrics indicate that the Whisper model effectively transcribes spoken language into text with high fidelity. The model's robust performance across diverse linguistic inputs demonstrates its versatility and effectiveness in handling transcription tasks for less-resourced languages. The Whisper model's success can be attributed to its ability to process a wide range of accents and dialects, making it a reliable choice for transcribing ethnic languages into Bangla.

5.5 Summary

The findings and examination of this chapter emphasize that the Transformer and

Whisper models are the most resilient and dependable options for translation and transcribing jobs, respectively. These findings establish a strong basis for creating effective and precise speech-to-text and translation systems for the specific ethnic languages into Bangla. The thorough assessment and comparative examination provide useful insights into the strengths and weaknesses of each model, informing future enhancements and applications in this domain.

CHAPTER 6
IMPACT ON SOCIETY, ENVIRONMENT AND SUSTAINABILITY

6.1 Impact on Life

The creation of a strong speech-to-text and translation system for ethnic languages into Bangla has significant consequences for individual lives. This technology enhances communication for those who speak minority languages, so encouraging linguistic inclusion and the preservation of cultural heritage. By granting persons access to fundamental services such as education, healthcare, and government services in their mother tongues, they can actively engage in society, thereby enhancing their standard of living. Moreover, this approach enhances individuals' agency by safeguarding their linguistic history, enabling them to preserve a robust bond with their cultural identity.

6.2 Impact on Society & Environment

Society:

The integration of this speech-to-text and translation system can significantly enhance societal cohesion by bridging language gaps. It fosters greater understanding and cooperation among different linguistic groups, promoting social harmony. Additionally, the system supports educational initiatives by providing learning materials in multiple languages, thus contributing to higher literacy rates and educational attainment among minority language speakers.

Environment:

From an environmental perspective, the deployment of this technology must consider the sustainability of resources used in its implementation. Cloud-based solutions for storing and processing linguistic data can lead to substantial energy consumption. However, optimizing the efficiency of these processes and using green energy sources can mitigate the environmental impact. Moreover, the system can reduce the

need for physical documentation and printed materials, contributing to lower paper consumption and waste generation.

6.3 Ethical Aspects

The development and deployment of this speech-to-text and translation system must adhere to ethical standards to ensure fair and responsible use. Key ethical considerations include:

Privacy and Data Security: Ensuring that the data collected from individuals is stored securely and used responsibly is paramount. Users must provide · informed consent, and their data should be anonymized to protect their privacy.

Bias and Fairness: The system must be trained on diverse datasets to prevent biases against any linguistic group. It is essential to ensure that the translation outputs are accurate and respectful of all cultural contexts.

Accessibility: The technology should be made accessible to all, regardless of socioeconomic status, ensuring that even the most marginalized communities benefit from it.

6.4 Sustainability Plan

To guarantee the enduring viability of the speech-to-text and translation system, the following procedures can be enacted:

Continuous improvement involves making regular updates and enhancements to the system, taking into account user feedback and technical changes. This practice helps to maintain the system's relevance and efficacy over time.

Community Involvement: Actively engaging with the communities that utilize the system can assist in upholding its cultural and linguistic precision. Engaging native speakers during the development and testing stages might yield significant perspectives.

Resource optimization: It involves the use of energy-efficient algorithms and renewable energy sources to reduce the environmental impact of data processing and storage.

Funding and Support: Acquiring financial resources from governmental entities, non-governmental organizations, and private sector collaborations helps ensure the system's sustainability and growth.

6.5 Summary

This chapter examined the extensive influence of the speech-to-text and translation system on individuals, society, the environment, and ethical concerns. Technology improves communication, fosters linguistic inclusion, and safeguards cultural legacy, so greatly enhancing the quality of life for speakers of minority languages. The societal impact of this involves promoting social cohesion and facilitating educational programmers, while from an environmental perspective, it requires the utilization of sustainable resources. Responsible use of this technology requires careful consideration of ethical factors, including privacy, data security, and justice. An enduring sustainability strategy, which prioritizes ongoing enhancement, community engagement, efficient resource utilization, and financial backing, guarantees the long-term viability and beneficial influence of the system.

CHAPTER 7
CONCLUSION AND FUTURE WORK

7.1 Conclusions

The primary objective of this study was to create an extensive speech-to-text and translation system capable of converting five ethnic languages—Malo, Bawm, Santali, Marma, and Garo—into Bangla. The study showcased the efficacy of diverse neural network architectures, such as Bi-LSTM, CNN Seq2Seq, GRU Seq2Seq, Seq2Seq, and Transformer models, in effectively addressing intricate linguistic tasks. The Transformer model has proven to be the most resilient and precise for translation, attaining the highest levels of accuracy throughout training and validation, and showcasing exceptional generalization abilities. The Whisper model demonstrated exceptional performance in transcription tasks, delivering accurate and dependable transcriptions with minimal Word Error Rates, even when presented with a wide range of linguistic inputs. The research findings emphasize the superior accuracy and efficiency of the Transformer model, which surpassed competing models in both the training and validation stages. The GRU Seq2Seq model demonstrated strong performance in terms of accuracy and efficiency, while the CNN Seq2Seq model had challenges in both training and validation trials, suggesting its limits in managing the complexity of the translation problem. The evaluation measures, such as accuracy, loss, validation accuracy, and validation loss, offered a comprehensive assessment of the performance of each model, affirming the strength of the Transformer and Whisper models. The created system has a profound effect on individuals by enabling communication in their original languages, fostering social unity, and safeguarding cultural heritage. This method empowers individuals who speak minority languages by providing them with access to key services in their original languages. As a result, it enhances their involvement in society and improves their overall quality of life. In summary, this research has established a strong basis for the creation of efficient speech-to-text and translation systems for ethnic languages. This contributes to promoting linguistic inclusion and preserving cultural heritage.

7.2 Further Suggested Works

To build upon the findings of this research, several future directions can be explored:

Expansion to More Languages: Extend the system to include additional ethnic and regional languages, further promoting linguistic diversity and inclusivity.

Enhanced Data Collection: Increase the dataset size and diversity to improve model robustness and accuracy, particularly for less-resourced languages.

Advanced Model Architectures: Investigate and integrate newer model architectures, such as more advanced Transformer variants or hybrid models, to enhance performance.

Real-Time Translation: Develop and optimize the system for real-time translation and transcription applications, making it more practical for everyday use.

7.3 Limitations

Although this research has made considerable gains and contributions, it is important to realize several limitations:

Data Limitations: The dataset, while extensive, is constrained in terms of both its size and diversity. Increasing the size of the dataset could significantly improve the performance of the model.

Model Complexity: The complexity of the Transformer model, although advantageous for performance, presents difficulties in terms of processing resources and training duration.

Generalization to Unseen situations: Although the models shown good performance on the test data, their capacity to adapt to completely novel and unfamiliar linguistic situations may differ.

Resource needs: Training and deploying these models necessitate high-performance computing resources, which can restrict accessibility in situations with limited resources.

Reference:

[1] Wang, C., Tang, Y., Ma, X., Wu, A., Popuri, S., Okhonko, D., & Pino, J. (2020). Fairseq S2T: Fast speech-to-text modeling with fairseq. arXiv preprint arXiv:2010.05171.

[2] Nagdewani, S., & Jain, A. (2020). A review on methods for speech-to-text and text-to-speech conversion. International Research Journal of Engineering and Technology (IRJET), 7(05).

[3] Hasan, H. M., Islam, M. A., Hasan, M. T., Hasan, M. A., Rumman, S. I., & Shakib, M. N. (2020, August). A spell-checker integrated machine learning based solution for speech to text conversion. In 2020 third international conference on smart systems and inventive technology (ICSSIT) (pp. 1124-1130). IEEE.

[4] Ahmed, S., Sadeq, N., Shubha, S. S., Islam, M. N., Adnan, M. A., & Islam, M. Z. (2020, May). Preparation of bangla speech corpus from publicly available audio & text. In Proceedings of the Twelfth Language Resources and Evaluation Conference (pp. 6586-6592).

[5] Zhang, M., Zhou, Y., Zhao, L., & Li, H. (2021). Transfer learning from speech synthesis to voice conversion with non-parallel training data. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 29, 1290-1302.

[6] Patil, A., Joshi, V., Agrawal, P., & Mehta, R. (2023, January). Streaming Bilingual End-to-End ASR Model Using Attention Over Multiple Softmax. In 2022 IEEE Spoken Language Technology Workshop (SLT) (pp. 252-259). IEEE.

[7] Bell, P., Fainberg, J., Klejch, O., Li, J., Renals, S., & Swietojanski, P. (2020). Adaptation algorithms for neural network-based speech recognition: An overview. IEEE Open Journal of Signal Processing, 2, 33-66.

[8] Z. Chen and H. Yang, "Yi Language Speech Recognition using Deep Learning Methods," 2020 IEEE 4th Information Technology, Networking, Electronic and Automation Control Conference (ITNEC), Chongqing, China, 2020, pp. 1064-1068, doi: 10.1109/ITNEC48623.2020.9084771.

[9] Dey, S., & Dutta, J. (2020, November). A low footprint automatic speech recognition system for resource constrained edge devices. In Proceedings of the 2nd International Workshop on Challenges in Artificial Intelligence and Machine Learning for Internet of Things (pp. 48-54).

[10] Nugroho, K., & Noersasongko, E. (2022). Enhanced Indonesian ethnic speaker recognition using data augmentation deep neural network. Journal of King Saud University-Computer and Information Sciences, 34(7), 4375-4384.

[11] Hasan, H. M., & Islam, M. A. (2020, August). Emotion recognition from bengali speech using rnn modulation-based categorization. In 2020 third international conference on smart systems and inventive technology (ICSSIT) (pp. 1131-1136). IEEE.

[12] Nayak, S. K., Nayak, A. K., Mishra, S., & Mohanty, P. (2023). Deep Learning Approaches for Speech Command Recognition in a Low Resource KUI Language. International Journal of Intelligent Systems and Applications in Engineering, 11(2), 377-386.

[13] Williams, A., Demarco, A., & Borg, C. (2023). The Applicability of Wav2Vec2 and Whisper for Low-Resource Maltese ASR. In Proc. 2nd Annual Meeting of the ELRA/ISCA SIG on Under-resourced Languages (SIGUL 2023) (pp. 39-43).

[14] Jain, R., Barcovschi, A., Yiwere, M., Corcoran, P., & Cucu, H. (2023). Adaptation of Whisper models to child speech recognition. arXiv preprint arXiv:2307.13008.

[15] Radhakrishnan, S., Yang, C. H. H., Khan, S. A., Kumar, R., Kiani, N. A., Gomez-Cabrero, D., & Tegner, J. N. (2023). Whispering llama: A cross-modal generative error correction framework for speech recognition. arXiv preprint arXiv:2310.06434.

[16] Kodali, M., Kadiri, S., & Alku, P. (2023). Classification of vocal intensity category from speech using the wav2vec2 and whisper embeddings. Proceedings of Interspeech'23.

[17] Park, C., Seo, J., Lee, S., Lee, C., Moon, H., Eo, S., & Lim, H. S. (2021, August). BTS: Back TranScription for speech-to-text post-processor using text-to-speech-to-text. In Proceedings of the 8th Workshop on Asian Translation (WAT2021) (pp. 106-116).

[18] Subakan, C., Ravanelli, M., Cornell, S., Bronzi, M., & Zhong, J. (2021, June). Attention is all you need in speech separation. In ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 21-25). IEEE.

[19] Graphcore AI, available at <https://www.graphcore.ai/hs-fs/hubfs/New%20chart.jpg?width=940&height=672&name=New%20chart.jpg>, last accessed on July 1, 2024, at 3:00 PM.

Appendix A

Complex Engineering Problems (EP) and Complex Engineering Activities (EA) Analysis

Title: Adapting Whisper AI for Multilingual Speech-to-Text Conversion in Bangladeshi Ethnic Languages

Students ID: [202-15-14440]

Students ID: [202-15-14439]

CO Description for FYDP

CO	CO Descriptions	PO
CO1	Integrate recently gained and previously acquired knowledge to identify a real-life complex engineering problem for the Final Year Design Project	**PO1**
CO2	Analyze different aspects of the goals in designing a solution for the Final Year Design Project	**PO2**
CO3	Explore diverse problem domains through a literature review, delineate the issues, and establish the goals for the Final Year Design Project	**PO4**
CO4	Perform economic evaluation and cost estimation and employ suitable project management procedures throughout the development life cycle of the Final Year Design Project	**PO11**

Addressing of COs, Knowledge Profile (K), and Attainment of Complex Engineering Problems (EP):

SN	EP Definition	Attainment	CO	Justification (with Knowledge Profile)	References
1.	**EP1: Depth of Knowledge required**	Yes (must be Yes)	CO1, CO2, and CO3	My project applies **K3** and **K4** by using system design, optimization, and computational modeling to develop a speech recognition system, incorporating signal processing, machine learning, and neural networks.	**Page no: 16-17**
				Our project addresses **K5** by meticulously designing a speech recognition system for Bangladeshi ethnic languages and it require the integration of advanced technological solutions that address K**6**.	**Page no: 16-18**

				Our project addresses **K8** through an extensive review and analysis of current and foundational literature in the fields of speech recognition and natural language processing.	**Page no: 07-09**
2.	**EP2: Range of Conflicting Requirements**	Yes/No	CO2	We encountered challenges while collecting data from remote areas of Bangladesh. Different ethnic groups lives in different part of Bangladesh, reaching them was a huge challenge. .	**Page no: 21-24**
3.	**EP3: Depth of analysis required**	Yes/No	CO2	Addressing CEP and EP-3 involves selecting the most effective solution from multiple options for adapting the Whisper Large V3 model to recognize Bangladeshi ethnic languages.	**Page no: 18-21**

4.	**EP4: Familiarity of Issues**	Yes/No	CO3	Addressing CEP and EP-4, this project explores uncharted territory by applying Whisper Large V3 for speech recognition of Bangladeshi ethnic languages—a domain not previously tackled extensively within computer science and engineering (CSE).	**Page no: 12**
5.	**EP5: Extends of application codes**	Yes/No	CO3	Addressing CEP and EP-5 involves adhering to best practices and standards in the development of the speech recognition system. This project complies with engineering codes of practice by incorporating ethical considerations, data privacy norms, and accessibility standards.	**Page no: 16-17**

5.	**EP6: Extends of stakeholders involved and conflicting requirements**	Yes/No	CO3	Addressing CEP and EP-6, the project engages with a diverse array of stakeholders including linguistic experts, technologists, community representatives, and end-users from Bangladeshi ethnic groups.	**Page no: 14-16**
6.	**EP7: Interdependence**	Yes/No	CO3	Using CEP and EP-7 as a starting point, this project aims to create a speech recognition system for Bangladeshi ethnic languages. This includes a lot of different problems, such as gathering data in places with few resources, analyzing different dialects, and making the technology work with advanced AI models like Whisper Large V3.	**Page no: 1-3**

7.	-	Yes/No	CO4	Addressing CO4 involves a detailed economic evaluation and cost estimation for the Final Year Design Project (FYDP), encompassing all phases from initial research and development to implementation and testing.	**Page no:** **26 - 27**

ORIGINALITY REPORT

10%	7%	4%	4%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to Daffodil International University Student Paper	2%
2	Roopa B. S., C. Christlin Shanuja. "chapter 12 Neural Network Approach of Combating the Data Security Issues", IGI Global, 2024 Publication	1%
3	"Deep Learning Applications, Volume 2", Springer Science and Business Media LLC, 2021 Publication	1%
4	github.com Internet Source	<1%
5	researchrepository.universityofgalway.ie Internet Source	<1%
6	Houssem Lahiani, Mondher Frikha. "A Comparative Study of Two Deep learning Architectures for Gesture Recognition on ArSL2018 Dataset", Research Square Platform LLC, 2024 Publication	<1%