Bengali Text to International Phonetic Alphabet (IPA) Transcription: A Comprehensive BYT5 Based Approach

by

S M Sazzad Bin Kamal
17101529

A thesis submitted to the Department of Computer Science and Engineering
in partial fulfillment of the requirements for the degree of
B.Sc. in Computer Science

Department of Computer Science and Engineering
Brac University
October 2025

Declaration

It is hereby declared that

1. The thesis submitted is my/our own original work while completing degree at Brac University.
2. The thesis does not contain material previously published or written by a third party, except where this is appropriately cited through full and accurate referencing.
3. The thesis does not contain material which has been accepted, or submitted, for any other degree or diploma at a university or other institution.
4. We have acknowledged all main sources of help.

Student's Full Name & Signature:

S M Sazzad Bin Kamal
17101529

Approval

The thesis titled "Bengali Text to International Phonetic Alphabet (IPA) Transcription: A Comprehensive BYT5 Based Approach" submitted by

1. S M Sazzad Bin Kamal(17101529)

Of Summer, 2025 has been accepted as satisfactory in partial fulfillment of the requirement for the degree of B.Sc. in Computer Science on October 10, 2025.

Examining Committee:

Supervisor:
(Member)

Md. Sabbir Ahmed
Lecturer
Department of Computer Science and Engineering
Brac University

Thesis Coordinator:
(Member)

Md. Golam Rabiul Alam, PhD
Professor
Department of Computer Science and Engineering
Brac University

Head of Department:
(Chair)

Sadia Hamid Kazi, PhD
Chairperson and Associate Professor
Department of Computer Science and Engineering
Brac University

Abstract

The International Phonetic Alphabet (IPA) is an alphabetic system of phonetic notation based primarily on the Latin script. The IPA is widely used by lexicographers, foreign language students and teachers, linguists, speech-language pathologists, singers, actors, language contractors, and translators. In the rapidly evolving digital era, the use of the International Phonetic Alphabet for the Bangla language is crucial to maintaining linguistic precision, aiding language learning, fostering research, preserving linguistic diversity, and facilitating communication and speech therapy. It plays a vital role in various aspects of the study of the Bangla language and related fields. In the work, it is planned to train T5-based Large Language Model (LLM) architectures on the collected dataset to convert Bangla texts into IPA notation and compare the LLM results to find the best language model. Our work extends beyond training, delving into the exploration of potential enhancement avenues. Another aspect is to investigate the impact of Bangla numerals, abbreviations, English texts, and out-of-vocabulary words. We trained and evaluated three T5-based architectures—ByT5, UMT5, and mT5—on a merged dataset of over 50,000 Bangla sentences representing both standard and dialectal speech. Through these explorations, we observe a spectrum of outcomes, where some modifications result in tangible performance improvements, while others offer unique insights for future endeavors in the fields of Natural Language Processing (NLP) and linguistic preservation.

Keywords: IPA; NLP; Phonetic alphabet; Dataset; LLM; Model training; T5 model;

Acknowledgement

Firstly, all praise to the Great Almighty Allah, for whom our thesis has been completed without any major interruption.
Secondly, to our advisor, Md. Sabbir Ahmed sir for his kind support and advice in our work. He helped us whenever we needed help.
And finally, for our parents, without their extensive support, it may not be possible. With their kind support and prayer we are now on the verge of our graduation.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

The International Phonetic Alphabet (IPA) is a standardized system of phonetic notation devised by the International Phonetic Association in the late 19th century. It provides a universally accepted set of symbols to represent the sounds of spoken languages, enabling linguists, educators, language learners, and speech professionals to transcribe speech with precision and consistency. The IPA has become indispensable in linguistic analysis and pedagogy, as it facilitates accurate pronunciation modeling and cross-linguistic comparison of phonological systems. By offering a comprehensive and systematic representation of articulatory phonetics, it supports both descriptive and comparative studies across a vast array of language families.

In practical contexts, the IPA is widely employed in contemporary dictionaries and pronunciation guides for languages such as French and English, ensuring standardized articulation of words across regional and educational boundaries [1]. Furthermore, it serves as a foundational tool for the orthographic development of languages that lack a standardized writing system [2]. Beyond descriptive uses, the IPA has also proven critical in historical linguistics, where it aids in tracking phonological change and reconstructing proto-language forms.

While extensive research has leveraged the IPA for European languages, relatively little attention has been directed toward its application in Bangla (Bengali), a language spoken by over 230 million people. Early computational work in this domain, such as that of Ahmad et al. [4], utilized sequence-to-sequence architectures based on Gated Recurrent Units (GRUs) to model Bangla pronunciation. However, these approaches lacked the flexibility and scalability now offered by modern transformer-based models. As a result, there remains a significant gap in IPA transcription research for Bangla, especially in dialect-rich or low-resource contexts.

To address this research gap, the present work explores the application of transformer-based large language models (LLMs) to the task of Bangla text-to-IPA transcription. Specifically, we adopt T5, an encoder-decoder transformer architecture that has demonstrated effectiveness in a wide range of natural language processing tasks, including machine translation, summarization, and text classification [3]. The model is trained on a curated dataset of Bangla text and phonetic transcriptions, with particular emphasis on pre-processing, regional variation, and sequence alignment. To assess system performance, we employ Word Error Rate (WER) as the principal

evaluation metric, consistent with established practices in speech recognition research [9].

This work represents a timely contribution to the field, offering a scalable, dialect-aware approach to Bangla phonetic transcription that builds upon recent advances in LLMs and IPA modeling.

1.1 Problem Statement

The International Phonetic Alphabet (IPA) remains a foundational tool across linguistics, speech technology, and language education, providing a standardized and precise system for representing spoken language. Its widespread application in major world languages, such as English and French, has enabled robust phonetic modeling, speech synthesis, and pronunciation training tools. However, the adoption and computational application of IPA in Bangla (Bengali)—a language spoken by over 230 million people globally—remain significantly underdeveloped. This deficiency presents a critical challenge within both theoretical linguistics and applied natural language processing (NLP), particularly in the context of phonetic transcription and speech processing systems.

Bangla's phonological system presents several unique challenges that complicate IPA transcription. These include inconsistent phonetic behavior of numerals, abbreviations, and loanwords (especially from English), as well as a high prevalence of out-of-vocabulary (OOV) items. Traditional approaches, such as pronunciation modeling using Gated Recurrent Unit-based Recurrent Neural Networks (GRU-RNNs), have yielded initial progress [4], but they fall short of delivering scalable, high-accuracy systems required for real-world linguistic and educational applications.

The emergence of transformer-based large language models (LLMs), particularly the T5 architecture, offers promising avenues to address these challenges. Transformer models are well-suited to handle the contextual and syntactic complexity of natural language and have demonstrated superior performance in a variety of NLP tasks, including machine translation, question answering, and text generation [3]. Despite their success, such architectures have not yet been applied to the problem of Bangla-to-IPA transcription, marking a significant research gap.

This research seeks to bridge that gap by developing a transformer-based transcription system using a fine-tuned T5 model. The system is designed to process Bangla text, including complex constructs like numerals and code-switching instances, and convert it accurately into IPA representations. Key challenges to be addressed include effective preprocessing to manage multilingual and non-standardized inputs, handling a large proportion of OOV words through data augmentation and model tuning, and evaluating the transcription quality using rigorous metrics such as Word Error Rate (WER) [9].

By addressing these challenges, the study aims to establish a scalable, context-aware, and linguistically grounded approach to Bangla phonetic transcription. The

outcomes will contribute to the advancement of speech technologies for underrepresented languages and provide a blueprint for incorporating IPA into low-resource NLP contexts. Ultimately, this work supports broader goals of linguistic preservation, digital inclusion, and the expansion of phonetic tools to a wider range of linguistic communities.

1.2 Research Objectives

This research aims to train a new text-to-IPA model that can handle Bangla numerals, Bangla signs, and abbreviations, especially given the limited work that has been done in this field. The objectives of this research are the following.

Objective 1

- Merge the standard [8] and the dialectal [11] datasets and analyze the new dataset.

Objective 2

- Handle Bangla numerals, abbreviations, English texts, and out-of-vocabulary words using data pre-processing and analysis.

Objective 3

- Fine-tune ByT5, UMT5 and MT5 in our new dataset, optimizing learning rate, batch size, and dropout for better convergence.

Objective 4

- Measure how accurately each model transcribed Bengali into IPA using WER (Word Error Rate) as our evaluation metric.

Chapter 2

Literature Review

This review synthesizes foundational and contemporary research in phonetics, speech synthesis, and computational linguistics, with an emphasis on French and Bengali phonetics. Particular attention is given to the role of **large language models (LLMs)** and transformer-based systems in advancing phonetic transcription and speech processing in both high- and low-resource language contexts.

2.1 Historical and Theoretical Foundations of Phonetics

The historical grounding of phonetics, particularly in French, is exemplified in Andrieux-Reix's review of Jean-Marie Pierret's work [1]. Pierret's Phonétique Historique du Français offers an exhaustive account of the diachronic development of French phonemes, addressing both the articulatory and acoustic dimensions. Andrieux-Reix emphasizes Pierret's pedagogical clarity and the text's role in establishing theoretical models that remain foundational for phonetic instruction and analysis. Similarly, Lambert-Drache's Sur le bout de la langue [2] provides an applied framework for learners, combining descriptive phonetics with classroom-oriented exercises. Her approach reinforces the importance of linking phonological theory to real-world usage, laying the essential groundwork for further computational modeling.

2.2 Advances in Neural Networks, Transformers, and LLMs

The introduction of the *transformer architecture* by Vaswani et al. [3] marked a turning point in natural language processing (NLP), laying the groundwork for the development of modern *LLMs*. These models, powered by self-attention mechanisms, are capable of learning complex patterns across vast corpora and have significantly improved tasks such as machine translation, speech synthesis, and phoneme prediction.

Based on this, Xue et al. [6] proposed *ByT5*, a token-free LLM that processes text at the byte level. This model effectively addresses tokenization challenges, especially in multilingual and low-resource settings, including morphologically rich

languages such as Bengali. Similarly, Liu et al. [5] demonstrated a transformer-based multilingual grapheme-to-phoneme (G2P) converter, emphasizing the adaptability of LLMs in educational and multilingual speech systems.

2.3 Applications in Speech Synthesis and Phonetic Transcription

LLMs have also shown strong performance in the domain of speech synthesis. Ahmad et al. [4] introduced a sequence-to-sequence model specifically tailored for the modeling of Bangla pronunciation, providing compelling evidence of the benefits of deep learning for underrepresented languages. Expanding on this work, Hasan et al. [8] developed a character-level Bangla text-to-IPA transcription model using a transformer-based LLM combined with sequence alignment techniques. Their model achieves high transcription accuracy, reinforcing the suitability of LLMs for phonetic applications in complex linguistic environments.

2.4 Multilingual and Multi-Speaker Systems

In multilingual and multispeaker speech recognition systems, efficient evaluation remains a critical challenge. Von Neumann et al. [9] investigated various definitions and computational strategies for the Word Error Rate (WER), a key metric in speech system evaluation. Their work is essential for assessing the performance of LLM-based systems, particularly in tasks involving overlapping speech, dialectal variations are common in Bengali and other South Asian languages.

2.5 Bengali Phonetics and Regional Language Modeling

Recent efforts have focused on developing LLM-based phonetic models for Bengali, a language with rich dialectal variation. The *Bhashamul project* [11] exemplifies community-driven innovation by using Kaggle as a platform to train region-specific IPA transcription models for Bengali. This initiative not only increases accessibility to phonetic tools but also highlights the potential of open-source LLM training to preserve linguistic diversity.

In extensions, Ahmmed et al. [10] introduced a novel approach using district-guided tokens to enhance the transcription of regional Bengali dialects into IPA. This methodology reflects the growing trend of tailoring LLMs to regional linguistic characteristics, thereby improving inclusivity and performance in low-resource NLP applications.

The integration of LLMs, particularly transformer-based architectures like ByT5, into phonetic and speech processing tasks has significantly advanced the field. Although foundational linguistic studies provide essential theoretical insights, LLMs offer scalable and adaptable solutions for speech synthesis, phonetic transcription, and multilingual NLP. This is especially impactful for underrepresented languages such as Bengali, where LLM can bridge the gap between linguistic diversity and computational accessibility.

Chapter 3

Methodology

This study employs a transformer-based approach to develop a Bangla text-to-IPA transcription system. Specifically, we evaluate and compare three T5-based large language models (LLMs)—ByT5, UMT5, and MT5—to determine the most effective model for this task. Each model is fine-tuned on a curated and comprehensive dataset designed to reflect both standard and dialectal Bangla phonological variations. The methodology consists of the following components:

3.0.1 Dataset Collection and Preparation

Two datasets were utilized and merged for training: a standard Bangla dataset derived from formal and literary sources [8], and a dialectal dataset compiled from colloquial and regional sources [11]. The merger was performed carefully to preserve phonological diversity while eliminating duplication. This unified dataset provides a more representative linguistic distribution, ensuring that the resulting models can generalize well across different Bangla dialects.

The dataset includes aligned Bangla text and IPA transcriptions. The inclusion of elements such as numerals, abbreviations, and mixed-language content (especially embedded English words) was emphasized to reflect real-world language use. Out-of-vocabulary (OOV) words were identified and flagged for special handling during both preprocessing and evaluation.

3.0.2 Data Preprocessing

Preprocessing steps were designed to enhance model performance by ensuring input consistency and reducing linguistic ambiguity:

- **Text Linearization:** Text normalization or linearization was applied to ensure consistent encoding of Bangla characters, such as date, signs. Punctuation and diacritics were standardized, and english alphabets and numerals were removed.
- **Numeral and Abbreviation Handling:** Bangla numerals were converted to their phonetic equivalents following the conventions established in [11]. Abbreviations were expanded or transcribed using context-aware rules.

- **Tokenization:**
  - For ByT5, byte-level tokenization was used, which encodes the input at the character level rather than word level. This approach, introduced by Xue et al. [6], allows the model to generalize well on low-resource or morphologically rich languages like Bangla.
  - For UMT5 and MT5, SentencePiece tokenization was applied, which segments text into subwords. This approach, based on the multilingual T5 model [6], offers a balance between vocabulary coverage and generalization, making it suitable for morphologically diverse languages.

3.0.3 Model Selection and Training

The following models were selected for comparative analysis:

- **ByT5:** A byte-level model with no pre-defined vocabulary, enabling it to handle any script or character input. Its robustness to noisy text makes it a strong candidate for low-resource phonetic tasks.
- **UMT5:** A universally pre-trained multilingual T5 model optimized for performance across various low- and high-resource languages. UMT5 uses a cross-lingual tokenizer and benefits from universal transfer learning [7].
- **MT5:** The original multilingual T5 model pre-trained on a massive multilingual corpus using the mC4 dataset [6]. It supports over 100 languages and is capable of zero-shot cross-lingual transfer.

Each model was fine-tuned using an encoder-decoder sequence-to-sequence architecture. The encoder receives Bangla text, and the decoder generates the corresponding IPA sequence. Hyperparameters such as learning rate, batch size, and dropout rate were optimized using grid search to ensure stable convergence.

3.0.4 Transcription Workflow

The transcription pipeline varies slightly for each model:

- ByT5 directly transcribes character-level Bangla input into IPA symbols, leveraging its byte-level generalization and ability to model character-to-character transformations. This approach is particularly effective for non-standard orthographic variants and OOV words.
- UMT5 and MT5 use subword-level encoding, which allows the models to capture longer-range dependencies between graphemes and their phonetic representations. These models rely on a fixed multilingual vocabulary, which may limit their performance on truly unseen or dialectal inputs, but enables better context modeling across longer sequences.

3.1 Description of the Data

To develop a robust Bangla-to-IPA transcription system, we compiled and merged two complementary datasets: a standard Bangla dataset and a dialectal dataset encompassing regional variations. This merger aimed to ensure both linguistic diversity and phonetic comprehensiveness.

Standard Bangla Dataset

The first dataset, used previously for training, comprises 27,749 sentence-level examples. Among them, 8,724 examples were initially encoded in non-UTF-8 formats and required re-encoding during preprocessing. Each sample contains approximately 12.4 words on average.

The dataset includes:

- 75,907 unique tokens in the test set and 45,035 in the training set.
- A total of 23,431 out-of-vocabulary (OOV) words were identified across datasets.

Figure 3.1 shows the word count histogram of the first training dataset, while Figure 3.2 presents the distribution of sample lengths. The maximum sample length is approximately 175 words in training and 200 in testing.

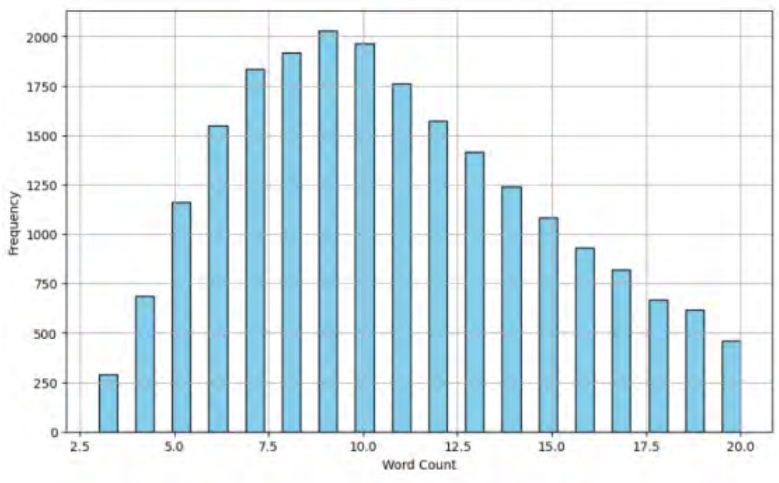

Figure 3.1: Word Count Distribution of the Standard Dataset Samples

Dialectal Dataset

The second dataset contributes regional dialects from six districts of Bangladesh: Rangpur, Kishoreganj, Narail, Chittagong, Nawabganj, and Tangail. This dataset consists of 30,031 sentence-level examples. Each region is fairly represented, ensuring a balanced dialectal spread.

Table 3.1 shows the number of samples per district.

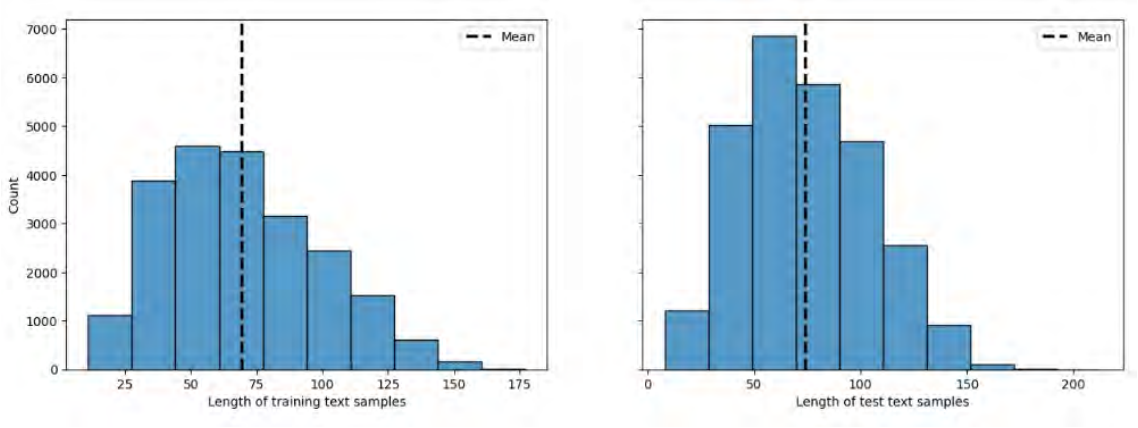

Figure 3.2: Length distribution of training and test samples

District	Number of Samples
Rangpur	5076
Kishoreganj	5061
Narail	5171
Chittagong	5022
Nawabganj	4994
Tangail	4707
Total	**30,031**

Table 3.1: Dialectal text samples across regions

The word distribution for this dialectal dataset is shown in Figure 3.3, highlighting the spread of vocabulary lengths across samples.

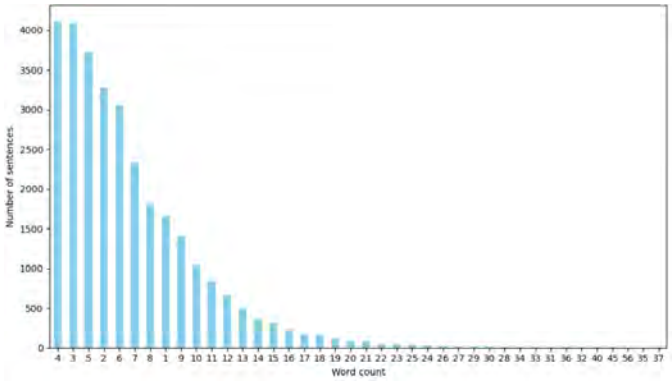

Figure 3.3: Word Count Distribution of the Dialectal Dataset Samples

Combined Dataset Summary

After initial pre-processing and cleaning of both datasets, they were merged into a single dataset to train the models (ByT5, UMT5, and MT5). This merged dataset captures both standard and dialectal usage, allowing the models to learn phonetic diversity and generalizable transcription patterns. Particular attention was paid to normalizing inconsistent encodings and addressing out-of-vocabulary issues.

3.2 Data Preprocessing

To ensure high-quality inputs for IPA transcription, a comprehensive pre-processing pipeline was implemented. This process addressed text normalization, encoding consistency, and structural transformation of both standard and dialectal Bangla data. The following key steps were taken:

- **Cleaning and Normalization:** All text samples were normalized by removing English alphabets, Latin-script characters, and Arabic numerals. This was done to maintain phonetic consistency, as such elements typically have inconsistent or context-sensitive pronunciations in Bangla.
- **Dataset Splitting:** From the annotated data, 90% was reserved for training and 10% for validation. The split was randomized with fixed seeds to ensure reproducibility across experiments.
- **Non-linear to Linear Text Conversion:** To make the data suitable for sequence transduction models, we transformed non-linear or symbolic structures (such as fractions, date formats, unit symbols, and abbreviations) into their spoken-word equivalents. This conversion ensured that the transcription models learn mappings from orthographic inputs to linear, phonetic forms. An example of this transformation is shown in Figure 3.4.

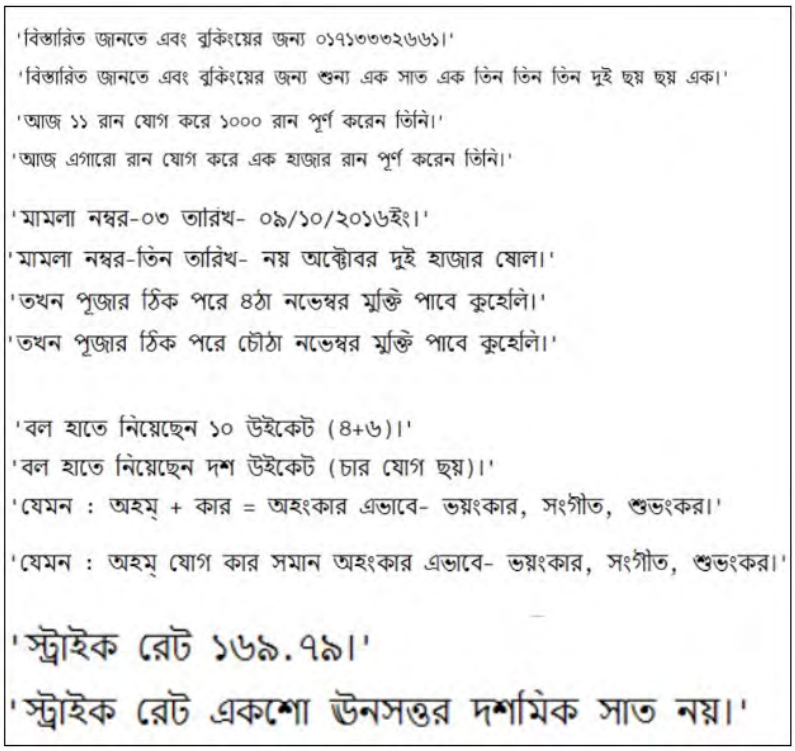

Figure 3.4: Non-linear to linear text conversion

This preprocessing pipeline not only improved model convergence but also reduced errors related to numerals, regional abbreviations, and OOV phenomena, enhancing the overall transcription accuracy.

3.3 Model Description

This study explores three transformer-based encoder-decoder architectures—ByT5, UMT5, and MT5—for the task of Bangla text-to-IPA transcription. Each model offers distinct architectural advantages, especially in multilingual and low-resource language settings, making them suitable candidates for evaluating IPA transcription effectiveness in Bangla.

3.3.1 ByT5

ByT5, introduced by Xue et al. [6], is a token-free variant of the T5 architecture that processes raw UTF-8 byte sequences directly. Unlike traditional subword-based models, ByT5 bypasses vocabulary-based tokenization by encoding input text as a sequence of 256 possible byte values. This design simplifies the preprocessing pipeline and enhances model robustness to text noise, spelling variation, and non-standard scripts—characteristics prevalent in Bangla data.

Key Features

- **Byte-Level Input Representation:** Each character is encoded as a byte, eliminating the need for tokenization or vocabulary construction.
- **Deeper Encoder Design:** The model adopts a deeper encoder relative to the decoder to manage longer byte sequences effectively.
- **Modified Pre-training Objective:** ByT5 employs a span corruption objective with longer average masked spans (20 bytes), suited for byte-level representations.

Advantages

- Strong robustness to noisy and irregular text.
- Competitive performance on generative and multilingual tasks.
- Simplifies preprocessing for low-resource or morphologically rich languages like Bangla.

Limitations

- Higher inference time due to longer input sequences at the byte level.
- Increased computational cost (FLOPs) compared to token-based models.

3.3.2 UMT5

UMT5, proposed by Chung et al. [7], extends the mT5 model to better support code-mixed, low-resource, and dialect-rich languages through universal pretraining and fine-tuning strategies. It introduces improvements in multilingual generalization by incorporating universal knowledge alignment and cross-lingual consistency objectives during training.

Key Features

- **Cross-Lingual Pretraining:** Improves representation learning for low-resource languages through shared multilingual embeddings.
- **Adaptability to Code-Mixed Inputs:** Well-suited for handling English–Bangla mixed text.
- **T5-Based Encoder-Decoder:** Retains the original encoder-decoder structure of T5 with language-aware modifications in pretraining data selection.

Application to IPA Transcription

UMT5 relies on subword tokenization (SentencePiece), requiring careful preprocessing to ensure accurate IPA alignment. Before training, text is normalized and segmented into subwords, then mapped to IPA symbols using supervised examples. This allows the model to generalize IPA conversion rules for both standard and dialectal Bangla.

3.3.3 mT5

The mT5 model, introduced by Xue et al. [6], is a multilingual version of T5 pretrained on the mC4 dataset, covering 101 languages. It uses the standard SentencePiece tokenizer with a shared vocabulary across languages, making it a strong baseline for multilingual tasks.

Key Features

- **Multilingual Pretraining:** Pretrained on a massive corpus (mC4) with cross-lingual data for wide language coverage.
- **Shared Vocabulary:** Uses a 250k token SentencePiece vocabulary across languages.
- **Flexible Transfer Learning:** Performs well in zero-shot and few-shot settings.

IPA Transcription with mT5

mT5, like UMT5, requires standard tokenization and preprocessing before fine-tuning. Bangla sentences are tokenized into subwords, and IPA labels are aligned at the sequence level. The model learns the mapping from Bangla subword tokens to IPA segments, benefiting from its multilingual generalization capabilities.

3.3.4 Model Summary

While mT5 and UMT5 leverage subword-based tokenization and strong multilingual generalization, ByT5 offers byte-level granularity and robustness against noise. Each model was fine-tuned on the merged Bangla dataset comprising both standard [8] and dialectal [11] pronunciations. Comparative evaluation using Word Error Rate (WER) provides a quantifiable measure of each model's transcription accuracy and ability to generalize phonetic rules in Bangla.

3.3.5 Comparative Performance on Multilingual Tasks

To assess the relative capabilities of ByT5 and mT5 across multilingual NLP tasks, we refer to the benchmark results provided in [6]. The evaluation encompasses a variety of tasks under three broad training paradigms: in-language multitask learning, translate-train, and cross-lingual zero-shot transfer. The models are evaluated on popular datasets such as WikiAnn NER, TyDiQA-GoldP, XNLI, PAWS-X, XQuAD, and MLQA.

Table 3.2 presents the performance comparison between mT5 and ByT5 across model sizes from Small to XXL.

Task	**Small**		**Base**		**Large**		**XL**		**XXL**	
	mT5	ByT5	mT5	ByT5	mT5	ByT5	mT5	ByT5	mT5	ByT5
In-language multitask										
WikiAnn NER	86.4	**90.6**	88.2	**91.6**	89.7	**91.8**	91.3	**92.6**	92.2	**93.7**
TyDiQA-GoldP	74.0 / 62.7	**82.6** / **73.5**	79.7 / 68.4	**86.4** / **78.0**	85.3 / 75.3	**87.1** / **79.2**	87.6 / 78.4	**88.0** / **79.3**	88.7 / 79.5	**89.4** / **81.4**
Translate-train										
XNLI	72.0	**76.6**	75.3	**79.9**	80.4	**83.6**	82.5	**85.3**	85.0	**87.1**
PAWS-X	83.8	**86.5**	87.5	**89.4**	90.4	**91.5**	91.6	**92.6**	92.5	**93.3**
XQuAD	58.6 / 45.2	**63.1** / **49.6**	68.3 / 54.7	**71.9** / **61.4**	73.9 / 56.7	**75.6** / **63.3**	75.4 / 61.1	**77.0** / **65.2**	77.1 / 61.4	**78.5** / **66.7**
MLQA	56.6 / 38.5	**61.7** / **45.9**	67.0 / 50.9	**70.1** / **56.3**	72.6 / 53.3	**74.2** / **60.1**	73.1 / 55.9	**74.8** / **60.5**	73.7 / 56.4	**75.2** / **61.5**
TyDiQA-GoldP	49.6 / 33.2	**57.5** / **39.5**	64.6 / 44.7	**69.1** / **51.2**	72.3 / 55.6	**75.5** / **62.3**	74.4 / 58.1	**76.7** / **63.6**	75.9 / 61.4	**77.9** / **65.9**
Cross-lingual zero-shot transfer										
XNLI	67.4	**69.4**	75.4	**75.8**	81.1	**82.0**	82.9	**83.6**	83.0	**83.7**
PAWS-X	82.5	**85.7**	87.5	**89.6**	90.6	**91.7**	91.8	**92.7**	92.4	**93.3**
WikiAnn NER	55.2	**63.7**	60.9	**66.6**	66.4	**68.4**	67.5	**68.7**	68.7	**69.5**
XQuAD	58.1 / 45.2	**63.7** / **49.7**	68.7 / 54.5	**71.5** / **60.1**	73.3 / 58.6	**75.5** / **63.3**	74.8 / 60.9	**76.8** / **65.0**	75.7 / 61.7	**77.3** / **66.9**
MLQA	56.6 / 38.5	**61.7** / **45.9**	67.3 / 51.7	**70.4** / **56.7**	72.7 / 54.4	**74.3** / **60.2**	73.1 / 56.6	**74.9** / **60.6**	73.7 / 57.2	**75.2** / **61.6**
TyDiQA-GoldP	36.4 / 24.4	**54.9** / **39.9**	54.0 / 37.7	**62.6** / **45.2**	68.4 / 50.7	**70.6** / **56.5**	72.4 / 55.8	**74.3** / **60.7**	72.8 / 57.1	**74.4** / **61.2**

Table 3.2: Performance comparison between mT5 and ByT5 on multilingual tasks. Bold values indicate higher performance between the two models. Results are taken from [6].

As seen in the table, ByT5 consistently outperforms mT5 across almost all tasks and model sizes. Notably, ByT5 exhibits significant gains in tasks involving spelling variation, code-switching, and morphologically rich languages, affirming its byte-level robustness. These improvements become even more prominent in zero-shot and low-resource scenarios, which are crucial for dialectal and phonetic applications like Bangla-to-IPA transcription.

Chapter 4

Evaluation and Metrics

In order to assess the performance and robustness of the Bengali text-to-IPA transcription models developed in this study, we adopted **Word Error Rate (WER)** as the principal evaluation metric. This chapter presents the motivation, formulation, and relevance of the chosen metric, along with a description of how it was applied to the outputs of the fine-tuned models—ByT5, UMT5, and mT5.

4.1 Evaluation Criteria

Transcription accuracy in grapheme-to-phoneme (G2P) or text-to-IPA tasks is generally evaluated using string similarity-based metrics. Among these, *Word Error Rate (WER)* is one of the most widely adopted measures due to its ability to capture discrepancies in predicted outputs at a word or symbol level. WER quantifies how many edits (insertions, deletions, or substitutions) are needed to convert the predicted transcription into the reference (ground truth) transcription.
WER is especially useful in evaluating the quality of output for speech systems, text normalization, and phonetic transcription, as it provides a holistic view of output accuracy without requiring manual inspection.

4.2 Word Error Rate (WER)

The *Word Error Rate (WER)* is defined as:

$$\text{WER} = \frac{S + D + I}{N} \tag{4.1}$$

Where:

- S = Number of substitutions
- D = Number of deletions
- I = Number of insertions
- N = Total number of reference words (ground truth)

A lower WER indicates better transcription performance. For example, a WER of 0.05 implies that, on average, there are five errors for every 100 IPA symbols predicted.

4.3 Evaluation Process

The models were evaluated using a curated Bangla test dataset, held out from training, that includes both standard and dialectal examples. The evaluation was performed using a custom inference pipeline built on top of the HuggingFace Transformers library. The pipeline executes the following steps for each input:

1. **Input Encoding:** The raw Bangla sentence is tokenized (or byte-encoded in the case of ByT5) and passed into the model.
2. **Prediction:** The model generates the predicted IPA sequence.
3. **WER Calculation:** The generated IPA is compared to the ground truth using WER, taking phoneme-level alignment into account.

Special care was taken to normalize the IPA outputs to avoid penalizing semantically equivalent variations (e.g., aspirated vs. unaspirated consonants in dialectal forms).

4.4 Why WER is Appropriate

WER is particularly well-suited to this study because:

- The output space (IPA symbols) is sequential and symbolic, like words in speech.
- It captures granular transcription errors, including minor phoneme-level mismatches.
- It enables fair comparison across models regardless of vocabulary size or architecture.
- It is consistent with prior literature in both speech recognition and phonetic transcription evaluation [9].

4.5 Limitations of WER

Although WER provides a reliable quantitative measure, it does not account for:

- **Phonological similarity:** Two IPA sequences may differ slightly but still represent very similar or acceptable pronunciations.
- **Multiple valid outputs:** In dialect-rich data, different IPA transcriptions might still be considered linguistically correct.
- **Alignment challenges:** WER assumes a fixed reference, which might disadvantage models producing reasonable variations.

To mitigate these limitations, qualitative inspection of outputs was performed alongside WER scoring (see Chapter 5), especially in cases where WER was high but outputs appeared phonetically reasonable.

4.6 Character Error Rate (CER)

To complement WER, we also report **Character Error Rate (CER)** as an evaluation metric. CER is widely used for sequence-to-sequence transcription tasks because it measures character-level accuracy, which is particularly important for morphologically rich languages like Bangla [12].

The CER is defined as:

$$\text{CER} = \frac{S + D + I}{N}$$

where S, D, and I denote substitutions, deletions, and insertions, respectively, and N is the total number of characters in the reference transcription.

We implemented CER using the `torchmetrics` library, applying the same procedure across ByT5, UMT5, and mT5 inference outputs. This ensures consistency in evaluation alongside WER.

This chapter has outlined the rationale behind using WER as the core evaluation metric for this research. By applying this metric across multiple T5-based architectures, we establish a consistent and interpretable way to compare model performance. The WER values reported in the next chapter offer insight into each model's capacity to generalize phonetic transcription rules across both standard and dialectal Bangla texts.

Chapter 5

Results and Discussion

In this chapter, we analyze and interpret the results obtained from the evaluation of three fine-tuned T5-based models—ByT5, UMT5, and mT5—on the Bengali text-to-IPA transcription task. Building on the evaluation procedures and metrics outlined in Chapter 4, we present both quantitative and qualitative findings, focusing on model comparison, dataset impact, and phonetic accuracy. We also explore key sources of error and the strengths and weaknesses of each approach.

5.1 Quantitative Results

The Word Error Rate (WER) scores of each model on the test dataset are summarized in Table 5.1. The results clearly show that ByT5 outperforms both UMT5 and mT5 by a significant margin.

Model	**WER**	**CER**
ByT5	0.0366	0.0096
UMT5	0.0545	0.0129
mT5	0.3136	0.0566

Table 5.1: WER and CER on the Bangla test dataset for different models

In addition to WER, we report CER scores. The CER values are consistent with WER trends, confirming that ByT5 achieves the lowest error rates, while UMT5 outperforms mT5 at both word and character levels. The inclusion of CER provides a finer-grained view of transcription accuracy, capturing character-level deviations that may not significantly affect WER.The inclusion of CER provides a finer-grained view of transcription accuracy, capturing character-level deviations that may not significantly affect WER [12].

The low WER of ByT5 demonstrates the effectiveness of byte-level encoding for handling morphologically rich, noisy, and dialectal Bangla data. In contrast, UMT5 and mT5, which rely on subword tokenization, struggle more with out-of-vocabulary items and dialectal variance.

5.2 Qualitative Analysis of Output IPA

To better understand model behavior, we examined randomly selected transcriptions across various linguistic contexts. Table 5.2 presents examples of IPA outputs and corresponding WER and CER scores for a range of sentence types. In many cases, ByT5 not only achieved lower error rates but also demonstrated better phonological accuracy—especially in handling aspirated consonants, compound words, and dialectal expressions.

The inclusion of CER highlights subtle character-level deviations that may not always contribute to WER, but still reflect transcription fidelity. For instance, dialectal variations often resulted in minor character substitutions that increased CER without significantly affecting WER. This makes CER a valuable complementary metric for evaluating fine-grained transcription accuracy in morphologically rich languages like Bangla.

Sentence	Ref. IPA	ByT5 IPA	WER	CER	MT5 IPA	WER	CER	UMT5 IPA	WER	CER
ইসরাইলের সাবেক লিকুদ পার্টির সরকারের রাজনৈতিক উপদেষ্টা ছিলেন সাফাদি।	ʃɪʃɾɐɪlɐr ʃɐbɛk lɪkuˌd pɐrtɪr ʃɔrkɐrɐr ɾɐɲoɪ̯tɪk upoˌdɐʃtɐ cʰɪlɐn ʃɐpʰɐˌdɪ.	ɪʃɾɐɪlɐr ʃɐbɛk lɪkuˌd pɐrtɪr ʃɔrkɐrɐr ɾɐɲoɪ̯tɪk upoˌdɐʃtɐ cʰɪlɐn sɐpʰɐˌdɪ.	0.1111	0.0389	ɪʃɾɐɪlɐr ʃɐbɛk lɪkuˌd pɐrtɪr ʃɔrkɐrɐr ɾɐɲoɪ̯tɪk upoˌdɐʃtɐ chɪlɐn sɐphɐˌdɪ.	0.2222	0.0389	ɪʃɾɐɪlɐr ʃɐbɛk lɪkuˌd pɐrtɪr ʃɔrkɐrɐr ɾɐɲoɪ̯tɪk upoˌdɐʃtɐ cʰɪlɐn sɐfɐˌdɪ.	0.1111	0.0519
তাহো খাবার লাগছে তো কেবল দুদ দিয়া ভাত নেয়সে আর দুদ দিয়া ভাতো খায়তো মেলা।	t̪ɐho kʰɐbɐr lɐgcʰe t̪o kebol ḍuḍ ḍɨʲɐ bʱɐt̪ neʲoʃe ɐr ḍuḍ ḍɨʲɐ bʱɐt̪o kʰɐʲt̪o mɛlɐ.	t̪ɐho kʰɐbɐr lɐgcʰe t̪o kebol ḍuḍ ḍɨʲɐ bʱɐt̪ neʲoʃe ɐr ḍuḍ ḍɨʲɐ bʱɐt̪o kʰɐʲt̪o mɛlɐ.	0	0	t̪ɐho khɐbɐr lɐgche t̪o kebol ḍuḍ ḍɨjɐ bfiɐt̪ nejoʃe ɐr ḍuḍ ḍɨjɐ bfiɐt̪o khɐjt̪o mɛlɐ.	0.5333	0.1	t̪ɐho kʰɐbɐr lɐgcʰe t̪o kebol ḍuḍ ḍɨʲɐ bʱɐt̪ neʲoʃe ɐr ḍuḍ ḍɨʲɐ bʱɐt̪o kʰɐʲt̪ mɛlɐ.	0.1333	0
বাংলাদেশ শত দরিদ্রতার মধ্য দিয়ে চললেও এর গণতান্ত্রিক কাঠামো ও গণতন্ত্রের প্রতি জনসাধারণের ভালোবাসা সত্যি বিরল।	bɐŋlɐḍeʃ ʃɔt̪o ḍorɪḍroʈɐr moddˌḍʱo ḍɨʲe colleo er gɔnoʈɐn-t̪rɪk ket̪ʰemo o gɔnoʈɔnt̪rer proʈɪ ɟɔnoʃɐḍʱɐroner bʱɐlobɐʃɐ ʃoʈʈɪ bɪrɔl.	bɐŋlɐḍeʃ ʃɔt̪o ḍorɪḍroʈɐr moddˌḍʱo ḍɨʲe colleo er gɔnoʈɐn-t̪rɪk ket̪ʰemo o gɔnoʈɔnt̪rer proʈɪ ɟɔnoʃɐḍʱɐroner bʱɐlobɐʃɐ ʃoʈʈɪ bɪrɔl.	0	0	bɐŋlɐḍeʃ ʃɔt̪o ḍorɪḍroʈɐr moddˌḍfio ḍɨje colleo er gɔnoʈɐn-t̪rɪk kɐthɐmo o gɔnoʈɔnt̪rer proʈɪ ɟɔnoʃɐḍfiɐroner bfiɐlobɐʃɐ ʃoʈʈɪ bɪrɔl.	0.3755	0.0428	bɐŋlɐḍeʃ ʃɔt̪o ḍorɪḍroʈɐr moddˌḍʱo ḍɨʲe colleo er gɔnoʈɐn-t̪rɪk ket̪ʰemo o gɔnoʈɔnt̪rer proʈɪ ɟɔnoʃɐḍfiɐroner bfiɐlobɐʃɐ ʃoʈʈɪ bɪrɔl.	0.6253	0.0071
আমার কাছ থেকে তার দায়িত্ব বুঝে নিতে হবে.	ɐmɐr kɐcʰ t̪ʰeke t̪ɐr ḍeʲɪʈʈo buɟʱe nɪ̀t̪e hɔbe.	ɐmɐr kɐcʰ t̪ʰeke t̪ɐr ḍeʲɪʈʈo buɟʱe nɪ̀t̪e hɔbe.	0	0	ɐmɐr kɐch t̪heke t̪ɐr ḍejɪʈʈo buɟfie nɪ̀t̪e hɔbe.	0.5	0.08	ɐmɐr kɐcʰ t̪ʰeke t̪ɐr ḍeʲɪʈʈo buɟʱe nɪ̀t̪e hɔbe.	0	0
ইন্টারনেট আর প্রযুক্তির অপব্যবহার কিভাবে মানুষের বিবেক আর মনুষত্বকে ধ্বংস করে দিচ্ছে সেটা দেখে ভীষণ মন খারাপ হয়.	ɪntɐrnet ɐr proɟukt̪ɪr ɔpobɛbohɐr kɪbʱɐbe mɐnuʃɐr bɪbek ɐr monuʃɔʈʈoke ḍʱɔŋʃo kore ḍɪccʰe ʃetɐ ḍekʰe bʱɪʃɔn mon kʰɐrɐp hɔ̀ẹ.	ɪntɐrnet ɐr proɟukt̪ɪr ɔpobɛbohɐr kɪbʱɐbe mɐnuʃɐr bɪbek ɐr monuʃɔʈʈoke ḍʱɔŋʃo kore ḍɪccʰe ʃetɐ ḍekʰe bʱɪʃɔn mon kʰɐrɐp hɔ̀ẹ.	0	0	ɪntɐrnet ɐr proɟukt̪ɪr ɔpobɛbohɐr kɪbfiɐbe mɐnuʃɐr bɪbek ɐr monuʃɔʈʈoke ḍfiɔŋʃo kore ḍɪcche ʃetɐ ḍekhe bfiɪʃɔn mon khɐrɐp hɔ̀ẹ.	0.3333	0.0465	ɪntɐrnet ɐr proɟukt̪ɪr ɔpobbɔbohɐr kɪbʱɐbe mɐnuʃɐr bɪbek ɐr monuʃɔʈʈoke ḍʱɔŋʃo kore ḍɪcche ʃetɐ ḍekhe bfiɪʃɔn mon khɐrɐp hɔ̀ẹ.	0.556	0.0078

Table 5.2: Results Comparison: Predicted IPA with WER and CER for Randomly Selected Sentences

5.3 Discussion

5.3.1 Impact of Dataset Diversity

The merged dataset, comprising both standard and dialectal Bangla, played a vital role in improving transcription robustness. The inclusion of regional language forms allowed the models—especially ByT5—to learn phonetic variation patterns. This diversity was particularly beneficial for dialectal transcription and OOV (out-

of-vocabulary) handling.

5.3.2 Model Architecture and Tokenization Effects

The results reaffirm the strength of byte-level architectures like ByT5 in low-resource and morphologically rich language settings. Unlike token-based models (UMT5 and mT5), ByT5 processes raw byte sequences, eliminating dependence on vocabulary segmentation and enabling better generalization to unseen patterns.

While UMT5 slightly outperformed mT5—likely due to better multilingual training—their subword representations still failed to fully accommodate the complexities of Bangla phonology.

5.3.3 Error Patterns and Limitations

Through detailed analysis, common error types were identified:

- **UMT5 and mT5:** Confusion between aspirated/unaspirated phonemes, and difficulty in handling numerals or abbreviations.
- **mT5 only:** Increased substitution and deletion errors in dialectal forms.
- **All models:** Occasional misalignment in rare words, especially code-mixed or out-of-domain content.

ByT5 was more resilient due to its span-masking pretraining and character-level processing but still made minor mistakes in very long or code-switched sentences.

5.4 Key Takeaways

- ByT5 significantly outperforms UMT5 and mT5 in Bengali IPA transcription, especially on dialectal and noisy inputs.
- Byte-level modeling offers greater flexibility for low-resource, morphologically complex languages like Bangla.
- Dataset quality and phonetic diversity are critical to achieving high transcription accuracy.
- Preprocessing steps like normalization and linearization directly influence performance by reducing ambiguity in inputs.
- CER complements WER by providing a more granular error analysis at the character level, which is especially relevant for morphologically complex Bangla.

This chapter provides a detailed understanding of the model behaviors and their limitations, preparing the ground for future enhancements discussed in the following chapter.

Chapter 6

Future Work

This study demonstrates the effectiveness of byte-level transformer models, particularly ByT5, for Bangla text-to-IPA transcription across standard and dialectal contexts. Despite promising results, several opportunities remain for improving accuracy, robustness, and real-world applicability. This chapter outlines key directions for future research and development.

6.1 Expansion of Dialectal and Domain Coverage

Although the current model was trained on six regional dialects, there remain unexplored variations in regions such as Sylhet, Barisal, and Noakhali. Including underrepresented sociolects and informal speech patterns would improve phonetic generalization. Additionally, expanding to domain-specific corpora (e.g., legal, medical, academic, and literary texts) could help the model handle specialized vocabulary and complex expressions.

6.2 Integration with Speech Systems

A natural extension of this work involves integration into downstream systems:

- **Text-to-Speech (TTS):** High-quality IPA transcriptions can serve as input for Bangla TTS engines, improving pronunciation and prosody in synthesized speech.
- **Automatic Speech Recognition (ASR):** IPA can be used as an intermediate representation to align predicted audio with true phonemes, enabling better error correction and model interpretability.

6.3 Post-processing with Phonological Rules

Although ByT5 performs well, some outputs contain minor systematic errors (e.g., handling of nasalization or compound consonants). These may be corrected using rule-based post-processing informed by formal Bangla phonology or tools like finite-state transducers (FSTs).

6.4 Improved Evaluation Metrics

Future work may explore integrating phonologically-aware evaluation metrics in addition to CER and WER.Metrics that consider phonetic similarity or multiple valid transcriptions could provide more linguistically meaningful evaluation [13].

6.5 Model Optimization for Deployment

For real-time or edge deployment, the model must be optimized for inference speed and memory usage. Possible strategies include:

- Model quantization or pruning
- Knowledge distillation into smaller student models
- Serving via lightweight ONNX or TensorRT pipelines

Such efforts will make the model viable for mobile apps, language learning platforms, and accessibility tools.

6.6 Cross-Language Generalization

The success of ByT5 suggests strong potential for applying this method to other low-resource languages with complex phonologies (e.g., Assamese, Odia, Sinhala). Using the same byte-level framework, transcription models can be developed to support digital inclusion in linguistically diverse regions.

6.7 Human-in-the-Loop Feedback Mechanisms

To maintain and improve performance after deployment, a human-in-the-loop setup can be implemented where users or linguists provide feedback on generated transcriptions. This feedback can be logged and periodically used to fine-tune or retrain the model.

6.8 Multimodal or Context-Aware IPA Generation

In future work, incorporating audio or semantic context could improve IPA predictions, especially for homographs or ambiguous words. Multimodal models that align acoustic cues with text could provide contextually richer IPA outputs.

Future research can significantly expand the applicability of this system through broader data coverage, integration with speech technologies, and human-guided fine-tuning. These improvements could ultimately enable high-fidelity transcription systems for education, accessibility, and NLP applications across Bangla and other under-resourced languages.

Chapter 7

Conclusion

This study addressed the underexplored problem of Bengali IPA transcription in a dialectal and OOV-heavy context. By comparing T5-based models, we demonstrate the viability of a byte-level transformer for high-accuracy phonetic transcription. Among the models evaluated—ByT5, UMT5, and mT5—the byte-level ByT5 model significantly outperformed the others in terms of Word Error Rate (WER), demonstrating its robustness in dealing with noisy, morphologically complex, and dialectally rich Bangla data.

The methodology involved merging standard and dialectal datasets, extensive data preprocessing including text linearization, and fine-tuning the models on this curated dataset. The comparative results confirm that ByT5's token-free, byte-level architecture offers superior generalizability, especially for low-resource and linguistically diverse scenarios. The strong performance of the model underscores the value of exploring non-tokenized approaches in underrepresented language processing.

Our findings contribute to the advancement of phonetic transcription and NLP tools for Bangla, paving the way for improved linguistic modeling, speech synthesis, and educational applications. In addition, the methodology and insights gained here can be extended to other low-resource languages, enhancing global linguistic inclusion.

In conclusion, this research not only addresses a key gap in Bangla phonetic modeling but also establishes a scalable, high-performing framework for future innovations in speech technology. With further refinement and deployment, such systems hold potential to significantly benefit areas such as text-to-speech (TTS), automatic speech recognition (ASR), language preservation, and regional accessibility in digital tools.

Bibliography

[1] N. Andrieux-Reix, "Pierret (jean-marie), phonétique historique du français et notions de phonétique générale, louvain-la-neuve, peeters, nouvelle édition, (série pédagogique de l'institut de linguistique de louvain, n° 19), 1994," *L'information grammaticale*, vol. 73, no. 1, 1997. [Online]. Available: https://www.persee.fr/doc/igram_0222-9838_1997_num_73_1_2934_t1_0070_0000_1

[2] M. Lambert-Drache, *Sur le bout de la langue: Introduction au phonétisme du français*. Canadian Scholars' Press, 1997, ISBN: 9781551301006. [Online]. Available: https://books.google.com.bd/books?id=isqDf8TKJLwC

[3] A. Vaswani et al., "Attention is all you need," in *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[4] A. Ahmad, M. M. Hussain, M. M. Selim, M. M. Iqbal, and M. S. Rahman, "A sequence-to-sequence pronunciation model for bangla speech synthesis," in *International Conference on Bangla Speech and Language Processing (ICBSLP)*, 2018, pp. 1–4. DOI: 10.1109/ICBSLP.2018.8554507

[5] J. Liu, C. Ren, Y. Luan, et al., "Transformer-based multilingual g2p converter for e-learning system," in *Artificial Intelligence in HCI*, H. Degen and S. Ntoa, Eds., Springer International Publishing, 2022, pp. 546–556, ISBN: 978-3-031-05643-7.

[6] L. Xue, A. Barua, N. Constant, et al., "Byt5: Towards a token-free future with pre-trained byte-to-byte models," *Transactions of the Association for Computational Linguistics*, vol. 10, pp. 291–306, 2022. DOI: 10.1162/tacl_a_00461 [Online]. Available: https://aclanthology.org/2022.tacl-1.17/

[7] H. W. Chung et al., "Unimax: Fairer and more effective language sampling for large-scale multilingual pretraining," *arXiv preprint*, 2023. eprint: 2304.09151. [Online]. Available: https://arxiv.org/abs/2304.09151

[8] J. Hasan, S. Datta, and A. Debnath, *Character-level bangla text-to-ipa transcription using transformer architecture with sequence alignment*, Preprint, Nov. 2023.

[9] T. von Neumann, C. Boeddeker, K. Kinoshita, M. Delcroix, and R. Haeb-Umbach, "On word error rate definitions and their efficient computation for multi-speaker speech recognition systems," in *ICASSP 2023 – 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5. DOI: 10.1109/ICASSP49357.2023.10094784

[10] S. Ahmmed, S. M. J. Islam, and S. H. Mustakim, *Transcribing bengali text with regional dialects to ipa using district guided tokens*, arXiv preprint, Mar. 2024. DOI: 10.48550/arXiv.2403.17407

[11] A. I. Humayun, Farigys, M. R. Hassan, et al., *Bhashamul: Bengali regional ipa transcription*, Kaggle competition, 2024. [Online]. Available: https://kaggle.com/competitions/regipa

[12] T. D. K, J. James, D. P. Gopinath, and M. A. K, "Advocating character error rate for multilingual asr evaluation," in *Findings of the Association for Computational Linguistics: NAACL 2025*, 2025, pp. 4926–4935.

[13] G. Wang, Z. Chen, B. Li, and H. Xu, "Cer-eval: Certifiable and cost-efficient evaluation framework for llms," *arXiv preprint arXiv:2505.03814*, 2025.