A character gram modeling approach towards Bengali Speech to Text with Regional Dialects

by

Md. Rezuwan Hassan
21266014

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

Department of Computer Science and Engineering
Brac University
November 2024

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:

Md. Rezuwan Hassan

21266014

Approval

The thesis titled "A character gram modeling approach towards Bengali Speech to Text with Regional Dialects" submitted by

1. Md. Rezuwan Hassan (21266014)

Of Fall, 2024 has been accepted as satisfactory in partial fulfillment of the requirement for the degree of M.Sc. in Computer Science on the November 3rd, 2024.

Examining Committee:

Supervisor:

(Member)

Dr. Md. Golam Rabiul Alam
Professor
Department of Computer Science and Engineering
BRAC University

Examiner:

(External)

Dr. Mashrur Imtiaz, PhD
Assistant Professor
Department of Linguistics
University of Dhaka

Examiner:

(Internal)

Dr. Farig Yousuf Sadeque
Associate Professor
Department of Computer Science and Engineering
BRAC University

Program Coordinator:

(Member)

Dr. Md Sadek Ferdous, PhD
Program Coordinator
Department of Computer Science and Engineering
BRAC University

Head of Department:

(Chair)

Dr. Sadia Hamid Kazi, PhD
Chairperson
Department of Computer Science and Engineering
BRAC University

Ethics Statement (Optional)

This is optional, if you don't have an ethics statement then omit this page
We hereby acknowledge that all of the co-authors of this work are aware of the provided ACL Code of Ethics and honor the code of conduct. All procedures performed involving human participants were by institutional and/or national ethical guidelines. All participants in the study provided informed consent. All the annotators and transcribers were paid above the local minimum wage (2,000 BDT were paid for each hour of transcription).

Abstract

The Bengali language, spoken in various regions of south-Asia and also among the Bengali diaspora, exhibits rich diversity with regional dialects or variations that reflect the cultural, geographic, and historical influences of different regional/socio-cultural communities. Based on phonology and pronunciation, Bengali is said to have 5 distinct major dialectal variations, such as Eastern Bengali Dialect, Manbhumi, Rangpuri, Varendri, and Rarhi. For the dialects present in Bangladesh, even finer stratification can be done based on the used vocabulary, pronunciation, phonology, syntax, and morphology.These regional Bengali dialects are found in regions such as Bangladesh in the regions of Chittagong, Sylhet, Rangpur, Rajshahi, Noakhali, Barishal, etc possess unique phonetic, lexical, and syntactic features that set them apart from standard Bengali and also unique from each other. However, research and resources dedicated to understanding and harnessing the potential of natural language processing of regional Bengali languages remain limited. To bridge this gap, this work aims to investigate and document the characteristics of regional Bengali languages through comprehensive data-driven linguistic analyses, including phonetic and morphological studies. We also aim to study the feasibility of developing computational models, including Automatic Speech Recognition (ASR) systems, tailored to regional Bengali languages, which can facilitate applications like virtual voice command assistants and language processing tools. Our research findings will contribute to the understanding of regional Bengali languages, paving the way to foster the advancement of language technologies that can cater to the diverse linguistic needs of Bengali-speaking communities. Through this study, we intend to promote preservation of the regional dialects of the Bengali language, foster cultural inclusivity, and facilitate effective communication in the Bengali-speaking regions.

Keywords: Automatic speech recognition; Regional Bengali speech; Wav2Vec2; Bengali Dialects; Linguistic Analysis; ASR; Dataset Curation

Acknowledgement

Firstly, I would like to thank my creator Allah, all praise to the Great creator. For making this journey possible for me and removing all the obstacles and major interruptions from the paths while completing my thesis.
Secondly, to my thesis supervisor, Professor Dr. Md. Golam Rabiul Alam sir for his support and advice during the thesis. He shared his wisdom and guidelines with me whenever I reached out to him with a query.
Also, a big thanks to Associate Professor Dr. Farig Yousuf Sadeque sir for playing a mentor role and for all the support, guidance, and wisdom.
Thirdly, the review panel of the conference "North American Chapter of the Association for Computational Linguistics". Though our paper was not accepted there, their reviews helped us greatly in our later works.

Next, to my parents, friends, and acquaintances. It wouldn't have been possible for me to continue this journey without their prayer and kind support.

A big thanks goes to team Bengali.AI, from team Bengali.AI I would personally want to thank Asif Shahriar Sushmit, Ahmed Imtiaz Humayun and Tahsin Reasat for their excellent guidance, Nazia Tasnim for helping with the modeling, and Trina Chakraborty for helping with few visualizations, Azmol Hossain, and Kanij Fatema for their linguistic support, Tanmoy Shome, Foriduzzaman Zihad, and Rubayet Sabbir Faruque for logistic support; and Siha Hoque for being a helping hand in when and where needed.

And finally, a big thank to all the assigned data collectors and transcribers shown in table 1 for their efforts.

Table 1: Name of all the data collectors and annotators

Rangpur	**Kishoreganj**	**Narail**	**Chittagong**	**Narsingdi**
Data Collectors				
Shaheen Alam	Mosharrof Hossain	Apon Sharif	Sajid Ullah Chowdhury	Monira Islam
Data Annotators				
Monisha Rani Mohonto	Antora dey	Md Apon Sharif	Ashikul Islam Shagor	Samad Mia
Md Rakibul Hasan	Mahjabin Rahman Maisha	S. M. Razoan Islam	Supratim Barua Koushik	M. M. Asif Yousuf
Nurun Nahar Sraboni	Rounake Afrose Eva	-	Nure Mehjabin	Tamima Islam Tonny

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

Language is fundamental for human expression and societal development, with structured language playing a pivotal role. In computer science, advancements in written and spoken language recognition have rapidly evolved. This evolution aims to enhance accessibility across various domains, including technology literacy and disability support [19].

Advancements in speech and writing recognition have spurred innovations in fields such as language education [38], language disorder assessment [21], and agricultural assistance [81]. Despite these advancements, the application of Speech-to-Text technology for regional variations of the Bengali language remains limited, largely as a result of the shortage of resources and available datasets.

Still, As per updated studies, there is a significant gap in both theoretical and computational linguistic studies focusing on regional Bengali variations, particularly in the context of automatic speech recognition tasks. Existing deep learning models have not adequately addressed these regional variations due to the lack of available datasets and resources [87]. Our research idea originated from this research gap in this domain and our research bridges these gaps by introducing an open-sourced dataset featuring diverse regional Bengali speech dialects and a modeling attempt that serves as a foundational resource for future research endeavors in this area.

Bengali, the fifth most spoken language globally [22], exhibits significant regional variations shaped by historical, geographical, and cultural influences. These dialectal differences encompass vocabulary, pronunciation, grammar, and cultural nuances, reflecting a sense of community identity and comfort among speakers.

1. **Vocabulary:** The standard Bengali language uses a consistent set of words that draw from both classical literature and modern usage. In contrast, re-

gional dialects often feature local expressions and slang that you might not encounter in formal or official contexts.

2. **Pronunciation:** How Bengali is spoken can vary significantly between the standard version and its regional dialects. Differences in how vowels and consonants are pronounced, as well as unique phonetic traits, can make each dialect distinct.

3. **Grammar:** The basic grammar of Bengali remains fairly uniform across different dialects, but there can be variations in how particles are used, verb forms are conjugated, and sentences are structured. Some dialects might simplify or modify certain grammatical rules.

4. **Cultural Influences:** Regional dialects often reflect local cultures and traditions, incorporating local sayings, idioms, and words borrowed from neighboring languages.

5. **Usage:** Standard Bengali is used in formal settings like education, media, and official documents. In contrast, regional dialects are more common in everyday conversations within communities and social groups, showing how people switch between different forms of the language based on the situation.

6. **Writing and Literature:** Formal writing, including newspapers, books, and academic work, typically uses standard Bengali. Regional dialects are rarely used in formal writing but might appear in casual communication like social media.

7. **Geographic Variation:** Different areas in Bangladesh and India have their own dialects, shaped by local cultures, histories, and landscapes.

Creating and implementing language policies that support standard Bengali while also respecting regional dialects is a challenging endeavor. This process requires a thoughtful approach, taking into account linguistic, cultural, and social factors. Language naturally evolves, and there can be resistance to changes in the standard form. Deciding which changes to accept or reject involves ongoing discussions and building consensus.

For an effective implementation model, it's important to recognize the rich linguistic diversity within Bengali. Many current datasets mainly focus on formal, standard

Bengali, which can limit users' ability to communicate in their natural style. This forces users to adapt to a more formal language, which can impede genuine communication. The distinct morphology and various accents within Bengali further complicate creating datasets that truly represent everyday language. Therefore, having large, diverse datasets is crucial for training comprehensive deep-learning models.

Progress in research and development has been hampered by a lack of resources, such as automatic speech-to-text systems for Bengali dialects, text-to-speech systems that include regional accents, and tools for dialect classification and transliteration projects. The absence of these resources presents major obstacles to advancing these technologies.

Accents are variations in pronunciation among speakers of the same language, while dialects have unique linguistic traits specific to certain groups or regions. These differences are shaped by geographic, cultural, and historical factors. For computational systems to effectively recognize and process these accents, it's essential to develop models that can accurately distinguish between them. This is crucial for tasks like speaker identification, emotion recognition, and understanding stress levels.

Alrehaili et al. [85] worked on analyzing Arabic dialects through a system that identifies dialects in audio recordings. They used a dataset of 672 audio samples from eight Arab dialects and applied Convolutional Neural Networks (CNN) for classification, achieving an accuracy of 83%. This work helps improve communication and translation across Arabic dialects.dialects.

Miao Wan et al. [83] focused on recognizing Chinese dialects using deep neural networks, with a special emphasis on regional accents. Their models achieved accuracies of 79.96% and 83.59%, enhancing applications like customer service and translation by addressing regional linguistic diversity.

In comparison, research on Bengali dialect recognition is still developing. Tomal et al. [82] are among the few to address this area. They focused on identifying regional Bengali dialects, which are primarily spoken and less documented. By analyzing extensive data from two dialects, Chatgaiya and Pabna, and using various data processing techniques, they achieved a high accuracy of 96%. Their work highlights the importance of preserving and understanding the diverse dialects within Bengali.

Despite significant efforts in compiling datasets, existing resources face limitations in speech diversity, size, and accessibility. To tackle challenges related to regional

languages, it is essential to develop dedicated resources tailored to these needs.

The SHRUTI read-speech corpus, published by IIT Kharagpur in 2011 [27], [28], [30], includes 21.64 hours of audio featuring 7383 distinct Bengali phrases and 49 phonemes. The IARPA voice corpus, part of the Babel initiative, contains 215 hours of Bengali phone conversations and scripted communications from 2011 and 2012, complemented by transcripts [37]. The Linguistic Data Consortium for Indian Languages (LDC-IL) acquired 138 hours of continuous Bengali speech in 2019 and subsequently developed a series of speech corpora, as noted in a 2020 research paper [57]. The Technological Development for Indian Languages Initiative (TDIL) [50] provides access to Bengali words from over $43,000$ audio recordings spoken by 1000 West Bengal native speakers. The European Language Resources Association (ELRA) offers a 70-hour Bengali speech corpus [44], although specifics such as publication date and corpus details remain undisclosed and unreviewed. A connected-word speech corpus [47] in 2018 was developed by Khan et al. comprising 62 hours of recordings from over 100 speakers . In the same year, Khan and Sobhan developed a second corpus focusing on isolated words, totaling 375 hours of recordings involving 150 speakers [46]. For the LVCSR challenge, Google released the "Large Bengali ASR training dataset" (LB-ASRTD) [48], a sizable Bengali speech corpus, in 2018 and made it accessible on the Open SLR website (Open SLR LB-ASRTD, 2018). It includes orthographic annotations in Bengali alphabets and consists of over 229 hours of conversation from 505 native Bangladeshis (323 men and 182 women), contributing a total of $217,902$ utterances.

In 2020, Ahmed and Sadeq [54] utilizing publicly available audio and text data, constructed a 960-hour annotated speech corpus and proposed a method for automatic transcription generation from existing audio recordings. SUBAK.KO [78] is a Bengali speech corpus from Bangladesh, comprising 241 hours of recordings from local speakers across eight divisions and thirty four districts of Bangladesh. Data were sourced from online platforms such as YouTube.
Developed by Bhogale et al. [77] in 2022, Shrutilipi ASR Corpus is a labeled ASR corpus derived from news broadcasts in 12 Indian languages, including Bengali. It includes nearly 6400 hours of data, with the Bengali segment comprising 443 hours. On 2023, To support and advance research in this specific domain, a significant initiative has been undertaken by Fazle et al. [87] and Bengali.AI resulting in the

creation of the Domain Diversified Bengali Speech Corpus named OOD-Speech: A Large Bengali Speech Recognition Dataset for Out-of-Distribution Benchmarking. It includes Bengali sentences over 2000 hours of transcribed audio recordings collected across India and Bangladesh. The initiative aims to compile 5000 hours of audio data by 2023, addressing significant linguistic challenges in Bengali speech recognition. This corpus constitutes an integral component of an ongoing initiative with the overarching goal of amassing a total of 5000 hours of meticulously collected audio data by the year 2023. This corpus delves into a comprehensive examination of the notable linguistic challenges entailed in the development of Bengali speech recognition systems. Moreover, we provide insightful analyses and observations derived from the wealth of data at our disposal.

Additionally, the study includes a benchmark analysis of the **Word Error Rate (WER)** for an ASR algorithm stemming from Hidden Markov Models and Gaussian Mixture Models (HMM-GMM). The research is dynamic, open to updates, and incorporates the latest benchmarks from the Common Voice Speech Corpus..

In summary, the key contributions of this research to the field include,

1. Pioneering the integration of a novel dimension in natural language processing development for the **Bengali** language. This initiative aims to enhance understanding and comprehension of **Bengali** from both computational and linguistic perspectives, catering to deep learning engineers/researchers and linguistic researchers.

2. Establishment of a comprehensive and standardized research pipeline accompanied by protocols designed to guide the entire research process. Tailored for similar research endeavors, this framework anticipates application across diverse Bengali-speaking regions in future studies.

3. Creation of a meticulously designed and curated 39-hour open-source speech corpus. This corpus is intricately crafted to capture nuanced regional speech patterns, encompassing various dialectical forms and variations, aimed at supporting research and development in the speech recognition domain and dialectal studies.

4. Development of a finely calibrated model capable of accurately transcribing regional Bengali speech. The model accommodates diverse dialects, variations in speech patterns, and unique vocabulary, while strictly adhering to established orthographic protocols.

5. Conducting a comprehensive linguistic analysis focusing on regional variations of the Bengali language across diverse geographical regions within Bangladesh. This analysis aims to provide insights into linguistic diversity and evolution within the Bengali language landscape.

The remainder of this paper is structured as follows:
Chapter 2 provides a comprehensive review of the existing research and development in this domain. Chapter 3 details the background study relevant to the topic. Chapter 4 outlines the methodology employed in the research. Chapter 5 presents the analysis of the results and Chapter 6 offers concluding remarks and summarizes the findings of the study.

Chapter 2

Related Work

Speech represents the most fundamental mode of human communication, uniquely free from literacy constraints. Advances in the speech recognition domain hold the potential to democratize global access to information and services. By facilitating seamless interaction with technology through spoken language, these advancements empower individuals of varying literacy levels and language proficiencies to effortlessly engage with digital platforms, access educational resources, and utilize essential services. Research in speech recognition has not only rendered previously inaccessible technologies accessible but has also enhanced task efficiency and user experiences fostering societal equity by narrowing digital divides and promoting broader participation in the digital era. [11].
This chapter examines existing research within the domain of speech recognition, focusing on available corpora and modeling techniques.

2.0.1 Speech Recognition Corpus

2.0.1.1 International speech corpora

Shivaprasad et al. [64] established a database for the Telugu language to facilitate a more comprehensive understanding of its dialects, addressing the absence of standardized resources for dialect study in speech recognition. They employed Hidden Markov Models (HMM) and Gaussian Mixture Models (GMM) to recognize these dialects based on speech patterns. Their findings indicated that the GMM technique outperformed HMM in this context.
MengEn Zhai et al. in their research [84], focused on the Yulin dialect, a lesser-studied regional dialect in China with inadequate data for speech recognition re-

search. Initially, they compiled and constructed a sole this specific dialect focused speech corpus, subsequently analyzing its phonological patterns and assembling a corresponding wordlist or dictionary. This methodology remarkably enhanced the Yulin dialect's speech recognition accuracy, achieving a 15.42% improvement over traditional methods in low-resource dialect scenarios.

2.0.1.2 Speech recognition modeling approaches for different languages

S. Darjaa et al. [43] implemented a basic computer program to differentiate between the accents of the Slovak language and its various regional dialects. They amassed a substantial collection of recordings from speakers with different accents, creating a specialized database to train the program for improved accent recognition. Additionally, they trained the system to recognize standard Slovak spoken in a neutral manner. The results indicated that the system could effectively identify the primary accent groups in Slovak, demonstrating its practical utility. However, the researchers suggest that with more data and advanced techniques, further improvements can be achieved.
Imaizumi, Ryo, and Masumura [60] developed a computer model to recognize different Japanese dialects. Due to limited data, they utilized both standard and dialect-specific data for model training. Initially, they mixed both datasets, but this approach resulted in the loss of dialect-specific details. Subsequently, they employed a supervised learning method, which reduced the error rate by 19.2%.

Swiss German presents significant challenges due to its diverse writing styles and lack of standardization. Nigmatulina et al. [63] investigated the comprehension of Swiss German across various dialects. They employed two approaches: dialectal writing (transcribing Swiss German phonetically) and normalized writing (transcribing Swiss German to resemble standard German). They tested these approaches using a computer program and a dataset encompassing 14 different Swiss German varieties. The normalized writing approach yielded better performance in terms of word error rates, while the dialectal writing approach excelled at recognizing individual letters. This research provides insights into improving algorithmic understanding of Swiss German and offers potential applications for other languages lacking standardized orthography.

Nguyen et al. [13] developed a system to recognize tonal variations in standard Vietnamese, particularly the Hanoi dialect. They employed wavelet transforms to analyze pitch variations in a large speech dataset. Their approach involved using Hidden Markov Models (HMMs) to recognize tones, finding that accuracy improved when using one of the two monotonic tones as a reference. They also developed an initial version of a system capable of understanding individual Vietnamese words, effectively incorporating the tonal recognition methodology.

At the 2020 International Conference on Intelligent Systems and Computer Vision (ISCV), Abderrahim Ezzine and his team presented a paper [58] on a system for recognizing speech in the Moroccan Darija dialect. They focused on recognizing the first ten Arabic digits spoken in Moroccan Darija. Their system utilized Hidden Markov Models (HMMs) along with Mel-frequency cepstral coefficients (MFCCs). Through various trials, they achieved a recognition accuracy of 96.27

Omar Aitoulghazi et al. [74] proposed an approach to develop an Automatic Speech Recognition (ASR) system for low-resource languages in their paper. They utilized Baidu's advanced "Deep Speech 2" model and tested it with 24 hours of spoken language data, achieving a Word Error Rate (WER) of 22.7% and a Character Error Rate (CER) of 6.03%. While these results indicate that the system is not yet perfect, they represent a promising step toward understanding the widely spoken Moroccan dialect.

Hamzah A. Alsayadi et al. [76] discussed a method to enhance ASR model performance for Arabic dialects. They developed a hybrid model combining a Convolutional Neural Network (CNN) and Long Short-Term Memory (LSTM) network with attention-based encoder-decoder techniques. Additionally, they created another language model using Recurrent Neural Networks (RNNs) and LSTMs, testing their approach on two sets of Arabic dialects. The experimental results showed that their methodology achieved a WER of 57.02

In a paper, Jooyoung Lee et al. [71] addressed the challenge of recognizing different dialects in the Korean language, which is complicated by a lack of research and resources. They proposed a novel approach focusing on intonation rather than individual syllables. They built a hybrid model combining spectral features such as fundamental frequency and trained it using a bidirectional LSTM (BiLSTM) network enabled with an attention mechanism. This approach helped identify dialect-rich parts of speech, even when mixed with non-dialect parts. Testing this model across

various ages and speech styles, they achieved an accuracy of 68.51

Vivek Bhardwaj et al. focused on improving ASR systems for Punjabi dialects in their research [69]. They used pitch acoustic features to enhance their system and tested it with two Punjabi dialects, Malwa and Majha. The results showed WERs of 23.25% and 25.91%, respectively, indicating that while there is room for improvement, the system represents progress in understanding these dialects.

In their research [40] , N. D. Londhe et al. employed two prominent Machine Learning (ML) techniques, Artificial Neural Networks (ANN) and Support Vector Machines (SVM), to develop an ASR system for the rare Indian dialect 'Chhattisgarhi'. They evaluated these methods using a dataset comprising 50 isolated words spoken by 15 individuals. The performance of ANN and SVM was compared to the traditional Hidden Markov Model (HMM). Additionally, they assessed the system's ability to recognize the same words spoken by different individuals. The results indicate that the ANN and SVM approaches are effective and reliable for recognizing this dialect.

2.0.1.3 Bengali speech corpora

In both Bangladesh and India, the speech recognition domain has attracted considerable research interest due to its profound impact on academic and industrial sectors. Over time, various datasets have been made available or actively utilized for training speech recognition models. These datasets are meticulously annotated at multiple levels, including phoneme, word, utterance, and sentence levels. They serve the dual purpose of facilitating model development and enhancing the understanding of the linguistic richness and diversity inherent in the languages spoken in these regions.

Due to the scarcity of publicly accessible research and development resources dedicated to this linguistic domain, Bengali, also known as Bangla, continues to be classified as a low-resource language [70].

To develop highly precise and effective models for various applications, the fundamental prerequisite is the availability of a substantial training dataset. However, in the context of Bengali speech recognition, the available resources are notably limited, presenting a significant challenge. While numerous datasets do indeed exist, it is noteworthy that the majority of the larger and potentially more comprehensive ones remain confined within private domains and have not been made publicly avail-

able. Remarkably, there is a conspicuous absence of datasets that encompass over a thousand hours of recorded Bengali speech that is sufficiently diverse in terms of linguistic content and context.

In their research article, Shafkat Kibria et al. [61] examine the impact of regional accents on the accuracy of ASR systems, emphasizing how pronunciation variations due to accents can impede ASR performance. The study focuses on two distinct groups of speakers, analyzing the acoustic characteristics of their accents regarding Bangladeshi Bengali, with particular attention to vowels such as monophthongal and diphthongal. The research includes both male and female speakers from the Sylhet region, known for its pronounced dialect differences, and individuals from other districts in Bangladesh with relatively milder dialect variations.

The analysis investigates acoustic features regarding accents such as pitch slope, formant frequencies, and vowel duration to classify and differentiate accents. The findings reveal significant differences in formant frequencies and pitch slope steepness among accents, which negatively impact ASR performance. The study highlights the necessity for accent focusing acoustic models to better accommodate speakers from regions with distinctive dialects. It also underscores the importance of incorporating accent-related speaker variability in corpora development to improve ASR systems for Bangladeshi Bengali.

For Bengali, there are limited speech corpora available, that too as open-resource. The SUBAK.KO speech corpus [78] stands out as a notable exception in the domain of linguistic resources. It is a publicly accessible annotated Bangladeshi standard Bangla speech corpus designed for research in automatic speech recognition. This corpus comprises 241 hours of exceptional quality speech data, consisting of 229 hours and 12 hours of read speech and broadcast speech data, respectively. The read speech recordings are in standard Bengali and were conducted in a controlled studio environment, featuring contributions from 33 male and 28 female, a total of 61 native Bangladeshi Bengali speakers, representing 8 divisions and 34 districts of Bangladesh. Additionally, the read speech section includes 1 hour and 30 minutes of recorded speech from two second language (L2) speakers. A portion of the dataset samples was obtained from popular online platforms such as YouTube and Facebook. This approach reflects contemporary language usage and ensures that the dataset aligns with modern communication practices. Each segment of SUBAK.KO has undergone meticulous manual annotation to ensure the accuracy and reliability

of the provided labels.

Table 2.1: Available Speech Corpus for Bengali. Stats of the datasets taken from [73] mostly.

Year	Corpus Name	Size of Dataset	No. of Speakers	Publicly Available
2011	SHRUTI [27], [28], [30]	21.64 hours	26 males, 8 females	Yes.
2012	IARPA-babel103b-v0.4b [37]	215 hours	Not known	Not Publicly Available. Access per application.
2014	LDC-IL [57]	138 hours	240 males, 236 females	No.
2014	TDIL [50]	43,000 audio files	1,000 native speakers	Not Publicly Available. Available for TDIL members.
2018	OpenSLR [48]	229 hours	323 males, 182 females	Not Publicly Available. Accessible under Attribution-ShareAlike 3.0 Unported (CC BY-SA 3.0 US)
2018	Bengali Connected Word Speech Corpus [47]	62 hours	50 males, 50 females	Not known
2018	Bengali Isolated Word Speech Corpus [46]	375 hours	50 males, 50 females	Not known
2018	OpenSLR [48]	229 hours	323 males, 182 females	Publicly Available under Attribution-ShareAlike 3.0 Unported (CC BY-SA 3.0 US)
2019	ELRA [44]	70 hours	Not known	Not Publicly Available. Available for ELRA members.
2020	Bengali Speech Corpus from Publicly Available Audio & Text [54]	960 hours	268 males, 251 females	No
2020	Subak.ko [78]	241 hours	33 males, 28 females	Yes
2022	Shrutilipi [77]	443 hours	All India Radio archives	Yes
2022	Common Voice Bengali Corpus [75]	1000 hours	22.1k Speakers	Not Publicly Available. Accessible under CC-0 Public Domain Attribution.
2023	OOD-Speech: A Large Bengali Speech Recognition for Out-of-Distribution Benchmarking [87]	2000 hours	25k Speakers	Yes

Among the limited publicly accessible resources, the most prominent is the OpenSLR dataset developed by Google [48]. However, this dataset primarily focuses on the utterance level (Table 2.1) of speech recognition, leaving a considerable gap in broader linguistic and contextual coverage.

Accents are variations in how words are pronounced, even among the same language speakers. It is noteworthy that accents mainly involve differences in pronunciation, whereas dialects encompass variations in pronunciation, vocabulary, and grammar. Thus, dialects represent a broader and more comprehensive linguistic concept, with accents constituting just one aspect of overall dialectal diversity.

Bengali, as a language, exhibits remarkable diversity with numerous distinct regional dialects. These dialects show noticeable variations in pronunciation and accent across different regions, including various districts in Bangladesh. In locations such as Dhaka, Chattogram, Sylhet, Noakhali, Barishal, and Rangpur, multiple dialects can be observed. Interestingly, even a single word can undergo pronunciation changes [93] depending on the specific region.

For developing an effective ASR system, the acquisition of an extensive and diversified training dataset that comprehensively represents the linguistic variations inherent in the target language is paramount. This dataset must be robust enough to enable the ASR system to accurately identify and transcribe speech, encompassing the rich tapestry of regional dialects and accents within the linguistic landscape.
Regrettably, a significant impediment exists on this path to ASR system excellence when considering the current state of available resources for the Bengali language. The prevailing corpus of Bengali speech data, although valuable, is notably limited in terms of its scale and scope. This constraint renders it incapable of adequately encompassing the entirety of the diverse dialectal nuances and regional linguistic variants that are integral to the Bengali language's linguistic diversity.
This inadequacy in available resources poses a significant challenge in constructing ASR systems that are both precise and reliable when dealing with the complexities of the Bengali language. This challenge becomes particularly pronounced in demanding ASR tasks, such as detecting spoken phrases within noisy acoustic environments or achieving speaker-independent speech recognition.
To overcome these impediments and enable the advancement of speech recognition technologies capable of effectively processing the spoken Bengali language, it is imperative to augment the existing resources. This augmentation should involve developing a more expansive, comprehensive, and representative Bengali speech corpus. These enhanced linguistic datasets would serve as the foundation for developing innovative ASR technologies, ensuring that the linguistic richness and diversity of Bengali can be effectively harnessed and understood by advanced speech recognition systems.

2.0.2 State-Of-The-Art (SOTA) Speech Recognition Systems

The emergence of modern deep learning technologies has catalyzed significant growth in the domain of speech recognition. Over the recent years, there has been substantial improvements in the speech recognition system's accuracy. Deep learning algorithms, particularly Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), have transfigured speech recognition task by enabling the development of more sophisticated models capable of learning from large-scale speech datasets.

Deep learning algorithms have enabled voice recognition systems to achieve unprecedented levels of speech recognition accuracy. These state-of-the-art models can now attain word recognition rates exceeding 95%, approaching human-like precision. This heightened accuracy has significantly improved the reliability of speech recognition systems and has opened up numerous new possibilities for the application of speech recognition technology.

A prime example of the advancement in speech recognition is the virtual assistants found on everyday devices. Google Assistant on Android, Siri on iOS, Amazon's Alexa, and Windows's Cortana are widely used to perform tasks such as note-taking, contacting individuals, or setting reminders. These virtual assistants are capable of perceiving voice commands and responding intuitively and spontaneously, thanks to the advanced development of voice recognition technology.

Other areas where speech recognition plays vital roles include recording and taking notes during meetings, translation and, transcription services, supporting video game interactions, and automating closed captioning for video indexing. Various companies offer such services through software or applications.

Automatic Speech Recognition is a research domain focusing on developing systems that can transcribe speech into text or other symbolic representations using models. ASR models are employed in numerous real-world applications, including powering voice-activated devices, facilitating language translation, and generating transcriptions for individuals with hearing impairments.

When constructing an ASR model, it is essential to consider several critical factors:

1. **Data collection and preparation:** ASR models require extensive audio datasets that accurately represent the target language and its regional dialects. These datasets must be transcribed and labeled for use in training and evaluation. Commonly used datasets include Common Voice [55], VoxCeleb [52], and LibriSpeech [36].

2. **Feature Extraction:** ASR models typically operate based on spectral features systematically derived from the raw audio waveform. These features enable the recognition and interpretation of spoken language with enhanced accuracy and efficiency.

3. **Acoustic Modeling:** This core component of an ASR system transforms audio information into textual output. Intricate statistical models map the extracted acoustic features from the input audio data to the corresponding textual representations. Deep Neural Networks (DNNs), known for their ability to model complex relationships within data, are frequently used for acoustic modeling in ASR systems.

4. **Language Modeling:** Language models predict the probability of different word combinations in the target language. They improves the precision and coherence of ASR output by infusing pre-existing linguistic knowledge into the recognition process. This involves methodologies such as n-gram models, which analyze word sequences up to 'n' words in length, and advanced neural network-based language models that leverage deep learning techniques to capture intricate linguistic patterns and relationships.

5. **Evaluation:** ASR models are typically evaluated using metrics such as WER or CER, which measure the degree of deviation from the reference transcripts. These metrics express the error rate as a percentage of incorrect words or characters.

Some of the top-performing models in the field include Whisper [79], Jasper [51], Wav2vec 2.0 [56], Conformer CTC [59], Kaldi ASR [31], and Google Assistant [25].

2.0.2.1 Wav2Vec 2.0:

Wav2Vec 2.0, developed by Facebook AI [56], represents a sophisticated ASR system. Its primary objective is to transcribe speech into written text accurately. The architecture is illustrated in Fig 2.1.

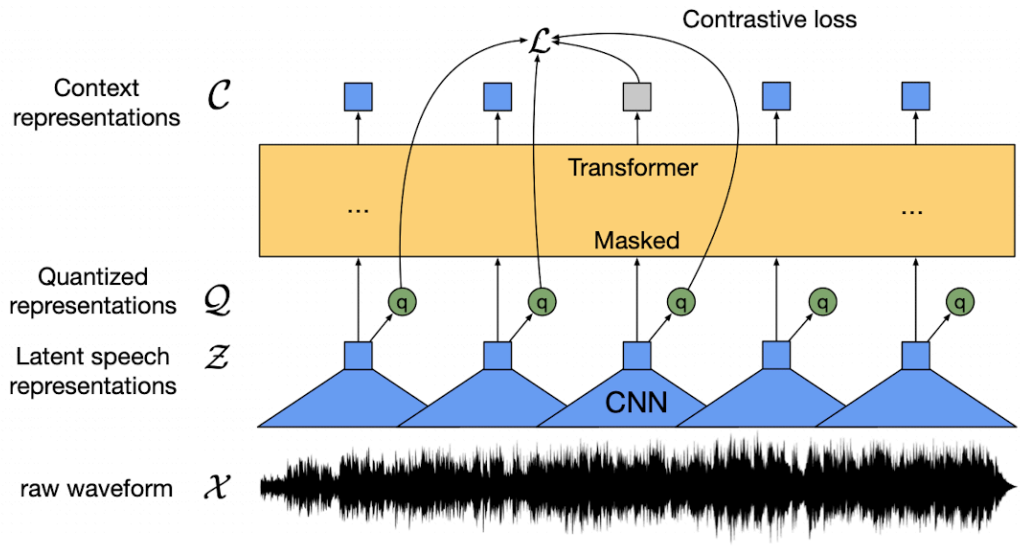

Figure 2.1: Illustration of the Wav2vec2 framework

A signature trait of Wav2Vec 2.0 is its innovative self-supervised learning approach, which enables the model to be trained without needing any transcribed speech data which makes the system particularly versatile, allowing it to be effectively applied to speech data with limited labelling.

Wav2Vec 2.0's core methodology requires a large corpus of unlabeled speech data to pre-train a neural network. Later a smaller, labeled dataset can be used to fine-tune this pre-trained network specific to a given ASR task. This approach has yielded exceptional performance across various speech recognition benchmarks.

Wav2Vec 2.0 marks a substantial improvement in ASR technology due to reducing the dependency on manual transcription efforts and improving the adaptability across different languages and domains. Its broad range of applications includes transcription services and voice assistants, making it an invaluable tool for the development of speech recognition systems.

What distinguishes Wav2Vec 2.0 is its self-supervised learning approach, which eliminates the need for transcribed audio data during training. This feature makes it particularly effective for languages and dialects with limited labeled data.

The foundational concept of Wav2Vec 2.0 involves pre-training a neural network

on a vast amount of unlabeled speech data. This pre-trained network is then fine-tuned with a smaller set of labeled speech data to perform specific ASR tasks. This methodology has demonstrated outstanding performance across various speech recognition benchmarks.

Wav2Vec 2.0 represents a significant advancement in ASR technology by reducing dependence on manual transcription efforts and enhancing adaptability across multiple languages and domains. Its applications are broad, ranging from transcription services to voice assistants, making it a valuable tool in the development of speech recognition systems.

The feature encoder output is discretized into a set of finite speech representations using product quantization [29], facilitating self-supervised training.

The Gumbel-Softmax technique facilitates the selection of discrete codebook entries in a fully differentiable manner [3], [34], [42]. The hard Gumbel-Softmax operations G [42] and the straight-through estimator [3] are employed to achieve this. The output of the feature encoder is projected onto $\mathrm{l} \in R^{G \times V}$ logits, and the equation below delineates the probabilities of selecting the v-th codebook entry for group g.

$$p_{g,v} = \frac{\exp(l_{g,v} + n_v)/\tau}{\sum_{k=1}^{V} \exp(l_{g,k} + n_k)/\tau} \tag{2.1}$$

In the equation 2.1, τ is a non-negative temperature, $n = \log(-\log(u))$ and u are uniform samples from $U(0, 1)$.

During the neural network training of the pre-trained model, codeword i is chosen by $i = \arg\max_j p_{g,j}$, and the true gradient of the Gumbel softmax outputs are used respectively in the forward and backward pass.

The pre-trained model has learned to represent the speech audio by optimizing two types of losses during its training, which are,

1. **Contrastive loss (L_m):** Distinguishes the correct quantized latent speech representation from a set of distractors to learn representations.

$$L_m = -\log \frac{\exp\left(\mathrm{sim}(c_t, q_t)/\kappa\right)}{\sum_{\tilde{q} \in Q_t} \exp\left(\mathrm{sim}(c_t, \tilde{q})/\kappa\right)} \tag{2.2}$$

in 2.2, c_t is the context network output centered over masked time step t, q_t is the true quantized latent speech representation in a set of $K + 1$ quantized candidate representations $\tilde{q} \in Q_t$ which incorporates K distractors along with

q_t and $\text{sim}(c_t, \tilde{q})$ is the cosine similarity between context representations c_t and quantized latent speech representations $\tilde{q}$.

2. **Diversity loss (L_d):** Promotes diversity in the learned representations and to utilize all the codebook entries uniformly.

$$L_d = \frac{1}{GV} \sum_{g=1}^{G} -H(\bar{p}_g) = \frac{1}{GV} \sum_{g=1}^{G} \sum_{v=1}^{V} \bar{p}_{g,v} \log \bar{p}_{g,v} \tag{2.3}$$

I equation 2.3, Each of the G codebooks has V entries. This is done by ensuring to utilization of the maximized entropy of the averaged softmax distribution l for each codebook entry $\bar{p}_g$ in an utterance batch. Note that the grumble noise is absent from the softmax distribution.
So, the total loss L becomes,

$$\mathcal{L} = L_m + \alpha L_d \tag{2.4}$$

I equation 2.4, α is a tuned hyperparameter to optimize the significance of the two losses.
Upon showcasing the extraordinary performance of Wav2Vec 2.0 on the widely recognized English ASR dataset such as LibriSpeech [36], Facebook AI launched a multilingual variant known as XLSR (cross-lingual speech representations). This model extends Wav2Vec 2.0's capabilities by enabling the acquisition of speech representations that are beneficial across multiple languages.
In November 2021, Arun Babu et al. released XLSR's successor, XLS-R (short for 'XLM-R for Speech') [68] which was pre-trained using audio data spanning 128 languages that was almost half a million hours in duration and the model is available in model sizes ranging from 300 million to 2 billion parameters.
Fine-tuning is done by adding a single layer on top of an existing pre-trained network and then train the model using own custom dataset and refine the model's performance on labeled data of audio downstream tasks like speech recognition/translation and audio classification as shown in the figure 2.2
XLS-R demonstrates significant outperformance over the existing state-of-the-art models in speech processing tasks like recognition/translation/diarization and language identification, as documented in the official report **XLS-R**.
Jasper: Developed by NVIDIA AI researchers [51], Jasper is an automatic speech recognition (ASR) system rooted in deep learning principles. Despite its modest

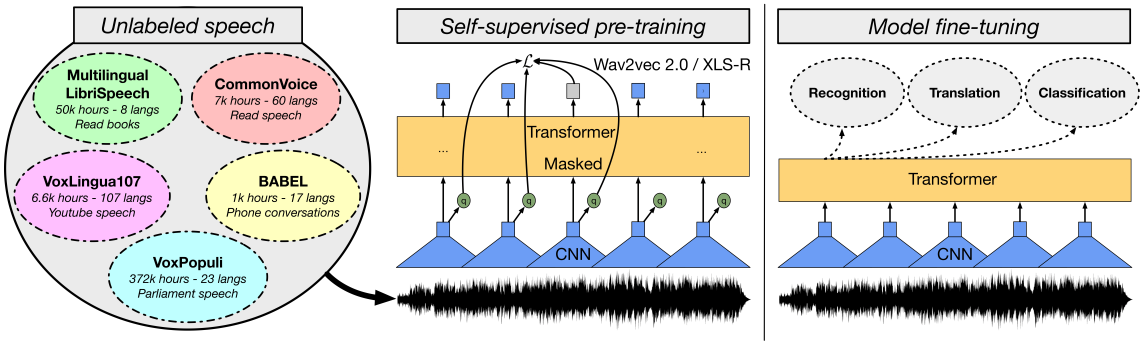

Figure 2.2: Wav2vec 2.0 / XLS-R

name, "Jasper" (Just Another Speech Recognizer) excels in converting spoken language into textual format with high efficiency.

A distinguishing feature of Jasper lies in its utilization of convolutional neural networks (CNNs) within its architecture. While recurrent neural networks (RNNs) have traditionally dominated ASR tasks, Jasper's adoption of CNNs enables parallel computation, facilitating faster audio data processing and potentially reducing computational overhead compared to conventional RNN-based ASR models.

Jasper has demonstrated competitive performance across diverse ASR benchmarks, underscoring its capability to accurately transcribe spoken language. Its versatile architecture supports adaptation to various languages and acoustic environments, enhancing its applicability across a broad spectrum of ASR applications.

To conclude, Jasper represents a significant advancement in the pursuit of efficient and precise ASR models. Recognized for its robust performance, Jasper contributes to ongoing advancements in speech recognition technology.

2.0.2.2 Conformer:

Conformer, as described by Gulati et al. [59], constitutes a deep learning framework tailored for automatic speech recognition (ASR) and natural language processing (NLP) tasks. It excels particularly in converting spoken language into textual form for ASR applications. Conformer architectures have garnered attention for their

adeptness in handling sequential data and achieving impressive performance across various ASR benchmarks.

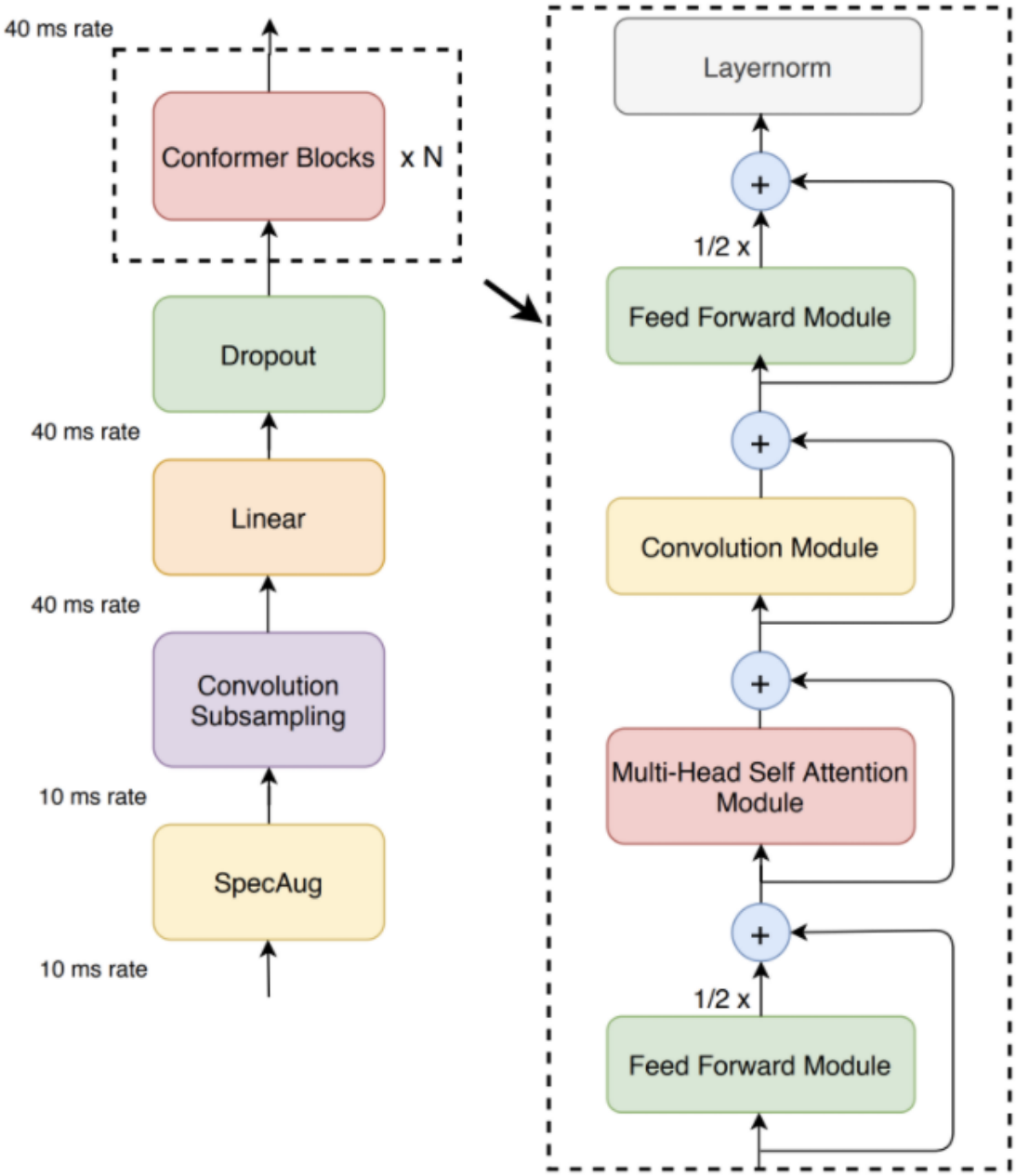

Figure 2.3: Conformer encoder model architecture

As depicted in Figure 2.3, the Conformer framework integrates several key components, including convolutional layers, a self-attention mechanism, and additional elements inspired by the Transformer architecture. The Conformer architecture consists of two macaron-like feed-forward layers that employ half-step residual connec-

tions, which encapsulate both multi-headed self-attention and convolution modules, followed by a post-layer normalization step.
The audio encoder first processes the input through a convolutional subsampling layer, subsequently passing it through multiple Conformer blocks, as illustrated in Figure 2.3. A distinctive feature of this model is the replacement of traditional Transformer blocks with Conformer blocks. Each Conformer block is structured with four stacked modules: a feed-forward module, a self-attention module, a convolution module, and a final feed-forward module.
This hybrid approach equips Conformer models with the capability to capture local and global dependencies within input audio data, thereby enhancing their efficacy in interpreting spoken language.

Key features of Conformer models include:

1. **Parallelization:** Conformer architectures are optimized for high parallelizability, facilitating accelerated training and inference processes. This attribute significantly enhances their efficiency in processing audio data.

2. **Self-Attention:** Leveraging a Transformer-inspired self-attention mechanism, Conformer models excel in capturing contextual dependencies across various segments of input sequences. This capability markedly improves their accuracy in speech recognition tasks.

3. **Depth and Stacking:** Conformer structures typically employ multiple layers stacked sequentially. This design enables them to effectively capture intricate patterns and features present in audio data, enhancing their robustness and performance.

4. **Adaptability:** By training on appropriate datasets, Conformer models can adapt seamlessly to diverse languages and dialects. This adaptability renders them versatile for a wide spectrum of ASR applications.

Conformer-based ASR models have consistently achieved state-of-the-art performance across numerous ASR benchmarks, underscoring their efficacy in comprehending and transcribing spoken language. Their adeptness in handling sequential data and their flexibility to accommodate various linguistic and environmental conditions have positioned them as a preferred choice for both academic research and

industrial applications in ASR.

2.0.2.3 Google ASR:

Google Automatic Speech Recognition (Google ASR) [25] is a robust speech recognition system developed by Google, integral to multiple Google services and products such as Google Assistant, Google Search, and Google Translate. This technology enables users to interact vocally with their devices, where Google ASR accurately transcribes spoken input into a textual format for subsequent analysis or actionable tasks.

Here are some key features and characteristics of Google ASR:

1. **Accuracy:** Google Automatic Speech Recognition (Google ASR) is renowned for its high precision in converting spoken language into text. It employs advanced machine learning algorithms and extensive datasets to continually enhance its recognition accuracy.

2. **Multilingual Support:** Google ASR offers robust support for a diverse array of languages and dialects globally, enabling users to interact with Google services in their preferred language.

3. **Voice Commands:** Google ASR drives voice-command functionalities, empowering users to manage devices, conduct web searches, schedule tasks, and perform various operations through spoken instructions.

4. **Voice Search:** Facilitating voice-based web searches, Google ASR enables convenient access to information and online resources without the need for manual typing.

5. **Accessibility:** Google ASR plays a pivotal role in accessibility initiatives, allowing individuals with disabilities to engage with technology through voice interactions.

6. **Natural Language Processing:** Integrated with sophisticated natural language processing (NLP) capabilities, Google ASR interprets spoken language in a conversational and context-aware manner, enhancing user interaction experiences.

7. **Cloud-Based Service:** Google ASR operates as a cloud-based service, offering developers the flexibility to integrate speech recognition functionalities into their applications and services.

8. **Privacy Considerations:** Google prioritizes user privacy and data security in its ASR implementations, ensuring users have control over their voice data and settings.

Google ASR continues to advance, fostering seamless and accessible voice interactions across a global user base. Its applications encompass a wide spectrum, encompassing voice assistants, search engines, transcription services, and voice-operated devices.

2.0.2.4 Whisper:

OpenAI's Whisper [79] currently stands as a discourse language model with significant potential that has garnered widespread attention. Similar to other models in the Whisper series, Whisper-medium comes pre-trained and is ready for use without the necessity for further fine-tuning. This makes it a practical option for numerous applications, including speech-driven interfaces and transcription services, where achieving a balance between speed and accuracy is essential.
Whisper has been meticulously trained and developed, exposed to a diverse dataset encompassing nearly 100 languages and over $680,000$ hours of carefully curated multilingual and multitask supervised data sourced from the web.

There exist two varieties of the model. English-only and multilingual. The English-only models focus on speech recognition, where the predicted transcribed text and the spoken audio are in the same language. Multilingual models, on the other hand, are trained for both speech recognition and speech translation. In speech translation, these models predict text transcriptions in a different language from the spoken audio. It has many varieties. The varieties and details are shown in the table 2.2
Similar to Wav2Vec 2.0, it performed fair on standard Bengali but performed very poorly on Regional data. This is because although it was trained on a large-scale dataset built using contents from the internet for Bengali, it was less than 1.5 hours of audio data. Notably, the training corpus lacks substantial representation of spon-

Table 2.2: Whisper model types details

Size	Parameters	English-only model	Multilingual model
tiny	39 M	Yes	Yes
base	74 M	Yes	Yes
small	244 M	Yes	Yes
medium	769 M	Yes	Yes
large	1,550 M	No	Yes

taneous Bengali speech with distinct regional accents and dialects, a gap addressed by the dataset introduced in this research.

As depicted in Figure 2.4, The Whisper model architecture, developed by OpenAI for automatic speech recognition (ASR) and language translation, is structured around a transformer-based encoder-decoder framework. The following are its key components:

1. **Encoder-Decoder Structure:** The architecture comprises an encoder and a decoder, facilitating efficient processing of audio input and the generation of text output.

2. **Audio Input Processing:** Raw audio signals are initially transformed into spectrograms, representing the frequency content over time, thus enabling structured analysis of the audio data.

3. **Multi-Head Self-Attention:** The model employs multi-head self-attention mechanisms, allowing it to attend to multiple segments of the input audio simultaneously. This capability is crucial for capturing long-range dependencies and contextual information.

4. **Positional Encoding:** Positional encodings are incorporated to maintain the sequential characteristics of audio data, providing the model with information regarding the temporal order of input frames.

5. **Feed-Forward Layers:** After the self-attention layers, feed-forward networks are utilized to further process the output, applying non-linear transformations that enhance the model's representational capacity.

6. **Layer Normalization:** Throughout the model, layer normalization is implemented to stabilize training and improve convergence rates.

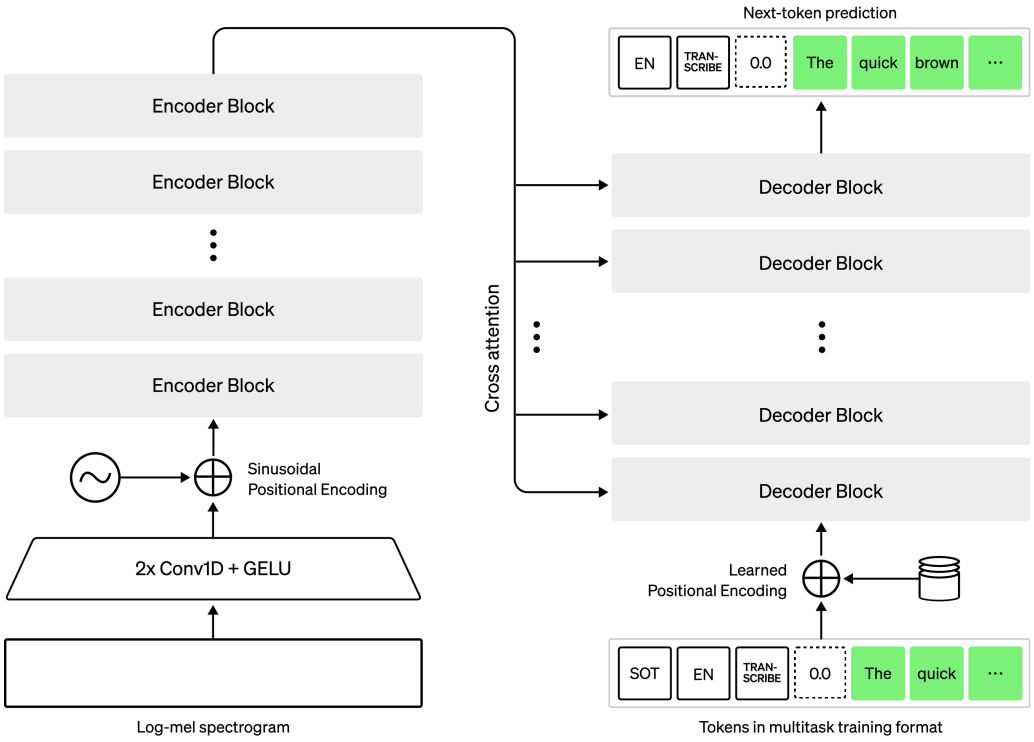

Figure 2.4: Conformer encoder model architecture

7. **Output Generation:** The decoder generates text sequences based on the encoded audio representation, utilizing both the attention mechanism and previously generated tokens to inform its predictions.

8. **Training Data:** The model is trained on a diverse and extensive dataset of multilingual speech, which significantly enhances its ability to handle various languages and accents.

In summary, the Whisper architecture leverages the strengths of transformer networks while being specifically optimized to address the challenges associated with speech recognition and translation tasks.

This work focuses specifically on augmenting the training dataset with regional Bengali accents and dialects, contributing significantly to ongoing research efforts. The dataset discussed herein highlights the manifold benefits of utilizing a comprehensive and diverse dataset, enhancing the model's robustness in navigating diverse accents, mitigating background noise interference, and understanding specialized technical

or regional languages such as Chittagong.

Beyond its contributions to speech recognition capabilities, Whisper's versatility extends to transcription and translation tasks across multiple languages and dialects, transcending linguistic barriers and promoting cross-cultural communication.

Moreover, Whisper's capability to expedite the translation of spoken content from various languages into English facilitates improved intercultural dialogue, fostering broader accessibility and understanding across diverse linguistic contexts.

Considering future developments, advancements in deep learning algorithms promise continued improvements in voice recognition accuracy, particularly in handling variations such as regional accents and dialects.

Chapter 3

Background Study

In this chapter, we will address the challenges associated with ASR modeling for regional dialects and provide a comprehensive linguistic analysis of each regional dialect.

3.1 Challenges with Bengali Speech Recognition With Regional Dialects

Bengali serves as the predominant language among the population of the Bengal region. Despite ongoing efforts towards standardization, regional dialects often prevail in everyday communication rather than the standardized form of Bengali. This trend is not exclusive to Bangladesh but mirrors similar patterns observed in other multilingual nations. Here, local dialects or languages are commonly used in daily interactions, contrasting with the use of standardized language reserved for formal settings such as education, media, and official documentation. As with many languages, variations in spoken language diverge significantly from the standardized norm in everyday contexts.

3.1.1 Deviation Challenges

The standard form of Bengali encounters several challenges, including the preservation of cultural identity. As languages evolve, there is concern that the process of standardizing Bengali may diminish certain cultural nuances and regional characteristics inherent in various dialects and local variations. Accessibility and inclusivity of the formal standard language pose challenges for certain demographic groups,

particularly those with lower levels of education, impacting their ability to comprehend and utilize it effectively. This disparity can lead to social exclusion and hinder access to education, information, and opportunities. Moreover, proficiency in the standard language may confer advantages in education, employment, and other domains, thereby exacerbating linguistic divides within society.

Another challenge involves the lack of modern terminology. Like all languages, Bengali grapples with the adaptation of new technological, scientific, and academic advancements. Developing and integrating new terminology while maintaining linguistic integrity presents a complex undertaking.

3.1.1.1 Regional Diversity

In the Bengali-speaking world, diverse local dialects, accents, and colloquialisms abound, shaped by geographical, historical, cultural, and linguistic interactions with neighboring regions. Informal settings, familial interactions, and casual conversations typically witness the use of these local or regional dialects. Such variations often incorporate distinct vocabulary, pronunciation, and grammatical structures that may diverge from the standardized Bengali form.

This natural evolution of language underscores the dynamic nature of linguistic usage and adaptation across different regions. While the standardized form of Bengali serves as a common medium of communication among speakers from diverse backgrounds, the proliferation of dialectal diversity enriches the language and reflects the cultural vibrancy of Bengali-speaking communities.

3.1.1.2 Cultural Diversity

Media outlets generally aim to uphold the standard language, yet regional influences can introduce variations in pronunciation and word usage, occasionally resulting in misunderstandings. Strict adherence to language standards may potentially constrain creative expressions, including poetry, literature, and other art forms that often derive impact from bending linguistic norms.

The prevalence of the standard language in official and formal contexts can sometimes evoke tensions among individuals who perceive it as emblematic of cultural and linguistic imperialism, particularly in regions marked by robust regional identities.

3.1.2 Linguistic Challenges

Language constitutes a dynamic entity, continuously evolving and adapting to cultural shifts, technological progressions, and interactions with other languages. Bengali, shaped by centuries of linguistic evolution, literary contributions, and cultural identity formation, holds a pivotal position in the history of the Bengal region. Its linguistic complexity is underscored by several distinctive features intrinsic to its structure. These include a sophisticated writing system, intricate inflectional morphology, the occurrence of gemination, and a rich inventory of diphthongs and triphthongs [14].
A detailed analysis is discussed in the later part of this chapter of this dissertation.

3.1.3 Dataset Acquisition Challenges

The SUBAK.KO speech corpus [78] stands out as the singular reported speech corpus to date, involving participation from 61 native speakers representing all eight divisions and thirty-four districts of Bangladesh. But the recordings were uniformly in standard Bangla and were executed under controlled conditions within a studio environment.

Research and development in this domain have been hindered by insufficient datasets and resources.

3.1.4 Modeling Challenges

Utilizing advanced state-of-the-art (SOTA) models for training poses significant computational challenges, particularly for academic researchers with limited access to such resources. It should be noted that achieving convergence during training is notably more difficult for Bengali compared to languages like English, primarily due to the greater diversity of distinct tokens within the Bengali language. Bengali, as an agglutinative language, exhibits complex morphological variations where individual words can manifest phonetic similarities yet differ significantly in their written forms.
Moreover, Bengali encompasses a wide array of regional accents and dialects, further complicating linguistic modeling efforts. While quantifying phonetic variations across these diverse dialects remains challenging, addressing this linguistic diversity

is crucial for effective modeling. Our approach involves comprehensive data collection that captures a broad spectrum of linguistic nuances in both text and audio formats. This meticulous curation aims to ensure the training corpus reflects diverse linguistic characteristics, providing neural networks with a robust dataset essential for effective ASR development in Bengali.

3.2 Linguistic Analysis

A detailed analysis from a linguistic aspect is discussed in this section.

3.2.1 Grapheme Diversity

The Bengali writing system is distinct because it employs orthographic syllables as graphemes, a characteristic shared with other alpha-syllabary or Abugida languages. Consequently, in this text, we use the terms *orthographic syllables* and *graphemes* interchangeably.

The Bengali language is written in the Bengali script, which is renowned for its astounding grapheme diversity [67]. It intricately combines consonant characters with built-in vowel sounds to form an *abugida* script. The Bengali Unicode system introduces multiple encoding methods for the same alphabet, necessitating normalization [24] and contributing to the challenges related to text grapheme diversity. Several of these appear in infrequent words and can also result from spelling errors.[65]. There are 11 vowel characters that seamlessly combine with consonants to form syllables and 39 basic consonant characters that represent various consonant sounds. The script modifies vowel sounds with diacritical marks, providing a variety of vowel combinations. It also includes numerals and punctuation marks in its repertoire and allows the construction of complex consonant clusters using conjunct characters. Occasionally, specific consonant combinations are represented by ligatures as well. Besides the standard vowel, consonant, and conjunct local dialect offers various vowel combinations in local dialect [66].

A significant quantity of consonant combinations poses difficulties in associating written characters with their spoken forms. This can hinder the effectiveness of Text-to-Speech (TTS) or Speech-to-Text (STT) models, as these combinations may have varying pronunciations depending on their context.

For example, সামনে is pronounced ʃɐmne whereas it is pronounced both the same and differently in different regions from where I have collected my data such as it is pronounced in the same manner in Kishoreganj and Narsingdi but সামনোত (ʃɐmnoṭo) in Rangpur, ছামনে ($c^hɐmne$) in Narail and সামনো (ʃɐmno) in Chittagong.

3.2.2 Inflection

In the field of linguistics, *inflection* refers to the process of creating different word forms by following a pattern derived from grammar, such as variations related to tense, or from the meaning of sentences, such as variations based on the gender of nouns. [9].

Similar to many other languages in the Indic language family, Bengali exhibits a high degree of inflection. Nouns, for instance, can manifest in more than 100 different case forms, while verbs can exhibit over 40 distinct variations. As a result, the count of whitespace-tokenized "words" in Bengali is notably high, reaching into the tens of millions, in stark contrast to the 188k words typically found in English [26]. Furthermore, Bengali's complex morphology allows for the creation of compound words, that results into the existance of substantial amount of Out of Vocabulary (OOV) words.

This issue poses considerable challenges for both TTS and ASR models, especially when dealing with languages with limited available resources, such as Bengali [49]. And when it comes to regional dialects, It exhibit unique inflectional patterns for gender, number, and case. For example, there are some distinctions in using singular and plural inflection in the Narsingdi and Kishoreganj dialects shown in the table 3.1

But the Chittagong dialect took some different inflection than the standard Bengali language shown in the table 3.2

Chittagonian dialects also exhibit unique inflectional patterns for gender variation shown in 3.3

Table 3.1: Morphological features of Narsingdi and Kishoreganj dialects

Standard Bengali	Inflection [IPA]	Narsingdi/Kishoreganj	Inflection [IPA]
আমটা [ɐmtɐ]	টা [tɐ]	আমডা [ɐmdɐ]	[dɐ]
আমগুলা [ɐmgulɐ]	গুলা [gulɐ]	আমডি [ɐmdɪ]	[dɪ]

Table 3.2: Morphological features of Chittagong dialects

Standard Bengali	Bengali Inflection [IPA]	Chittagong	Chittagong Inflection [IPA]
গরুগুলো [gɔrugulo]	গুলো [gulo]	গরুউন [gɔru͡u̯n]	উন [uu̯]
দাঁড়িগুলো [d̪̃ɐɪgulo]	গুলো [gulo]	দাঁইরউয়া [d̪̃ɐ͡ɪru̯ʲɐ]	উয়া [u̯ʲɐ]
গাছ থেকে [gɐcʱ tʱeke]	থেকে [tʱeke]	গাছত্তুন [gɐcʱo̯t̪̯tun]	তুন [t̪un]
বউয়ের [bo͡u̯er]	এর [er]	বউঅর [bo͡u̯ɔr]	অর [ɔr]

Table 3.3: Chittagong dialects inflectional patterns for gender variation.

In English	In Standard Bengali [IPA]	In Chittagonian [IPA]
He Does	সে করে [ʃe kɔre]	ইতে গরে [ɪt̪e gɔre]
She Does	সে করে [ʃe kɔre]	ইতি গরে [ɪt̪ɪ gore]

In the Rangpur region, This process includes combining the stem with a grammatical morpheme that does not produce new words but rather indicates aspects of the grammatical function of a word shown in the table 3.4
Another notable difference in the Rangpur dialect is observed in the inflections used in verb forms. The particular differentiation is outlined below in table 3.5

Table 3.4: Morphological features of Rangpur dialects

Standard Bengali (IPA)	Root word	Inflectional [IPA]	Rangpur [IPA]
অফিসে [ɔfɪse]	অপিস [ɔpɪs]	-এত [-et̪]	অপিসেত [ɔpɪset̪]
মানুষের [mɐnuʃer]	মাইনষ [mɐɪnʃ]	-এর [-er]	মাইনষের [mɐɪnʃer]
মানুষকে [mɐnuʃke]	মাইনষ [mɐɪnʃ]	-এক [-ek]	মাইনষেক [mɐɪnʃek]
বাজারে [bɐʃɐre]	বাজার [bɐʃɐr]	-ওত [ot̪]	বাজারোত [bɐʃɐrot̪]
ছিল [cʰɪlo]	ছিল [cʰɪlo]	আ [ɐ]	আছিল [ɐcʰɪl]
বিছানায় [bɪcʰɐnɐe]	বিছনা [bɪsnɐ]	-ত [t̪]	বিছনাত [bɪsnɐt̪]

Table 3.5: Inflectional Verb Forms in Rangpur Dialect

Standard Bengali [IPA]	Rangpur Dialect [IPA]
(আমি) খাব [ɐmɪ kʰɐbo]	(মুই) খাবু [muɪ kʰɐbo]
(আমি) খেয়েছি [ɐmɪ kʰeʲecʰɪ]	(মুই) খাইসি [muɪ kʰɐɪsɪ]
(কাজ) করবে না? [kɐj kɔrbe nɐ]	(কাম) করবান নয়? [kɐm kɔrbɐn no͡e̯]
(আমি) শুনলাম [ɐmɪ ʃunlɐm]	(মুই) শুননু [muɪ ʃununu]
দেখি [d̪ekʰɪ]	দেহং [d̪ehɔŋ]
করার [korɐr]	করং [kɔrɔŋ]

Table 3.6: Morphological features of Narail dialects

Standard Bengali [IPA]	Root word [IPA]	Inflectional [IPA]	Narail Dialects [IPA]
লোকজনকে [lokɟonke]	লোকজন [lokɟon]	রে [ɾe]	লোকজনরে [lokɟonɾe]
মেয়ের [meʲer]	মাই [mɐɪ]	এর [er]	মাইয়ের [mɐɪer]
দোকানটা [d̪okɐntɐ]	দুকান [d̪ukɐn]	ডা [dɐ]	দুকানডা [d̪ukɐndɐ]

3.2.3 Aspiration

Aspirated consonants খ, ঘ, ছ, ঝ, ঠ, ঢ, থ, ধ, ফ, ভ) [kʰ, gʱ, cʰ, ɟʱ, tʰ, dʱ, ʈʰ, ɖʱ, pʰ, bʱ] are consonants that are accompanied by secondary articulation. [10]. In Standard Bengali, aspiration is found in voiceless and voiced stops. Both voiced aspiration and voiceless aspiration occur in the language. Aspirated consonants create difficulties in both ASR (Automatic Speech Recognition) and TTS (Text-to-Speech) modeling because of their subtle pronunciation nuances. This complexity is further compounded when dealing with exaggerated aspiration in breathy voices, making it particularly challenging to model natural speech [23]. Additionally, individuals with pronounced accents and non-native speakers tend to pronounce aspirated sounds differently, adding to the challenges of achieving accurate generalization.[39].

It is common to find aspirated sounds /kʰ/, /tʰ/, /bʰ/, /gʰ/, /dʰ/, /pʰ/ are changed into non-aspirated /k/, /d/, /b/, /g/, /p/ sounds. Shown in the 3.7, the entry in the first row is a case of bilabial plosive where ʰ indicate a voiceless aspiration but the second and third entries are cases of voiced stop sounds where ʱ is indicating a voiced aspiration.

Rangpuri, alternatively known as Kamtapuri or Rajbanshi, represents an Eastern Indo-Aryan language spoken in specific regions of Bangladesh and India, particularly within

Table 3.7: Inflection example of standard Bengali

Voiceless aspiration [IPA]	English Translation	Voiced aspiration [IPA]	English Translation
ডাকা [dɐkɐ]	Call	ঢাকা [dʱɐkɐ]	Cover
দান [d̪ɐn]	Donate	ধান [d̪ʱɐn]	Paddy

Table 3.8: Aspiration in Rangpur Dialect

Rules	Standard Bengali [IPA]	Rangpur Dialect [IPA]
/kʰ/ - /kk/	দুঃখের [d̪uʃkʰer]	দুক্কোর [d̪ukkor]
/ʈʰ/ - /d/	ওঠে [oʈʰe]	ওডে [ode]
/bʰ/ - /b/	লাভ [labʱ]	লাব [lab]
/d̪ʱ/ - /d̪/	বন্ধু [bond̪ʱu]	বন্দু [bond̪u]

the Rangpur division of Bangladesh. While it shares specific phonological and lexical features with Standard Bengali, it maintains a close mutual intelligibility. The widespread influence of the standard Bengali has led individuals from other communities and regions to adopt it across various aspects of life. Therefore, this shift in the use of language has resulted in a decline in the usage of their native dialect. From my collected and transcripted data of the Rangpur dialect, the following non-aspiration sounds are found which are shown in 3.8

For the district Kishoreganj, the table 3.9 shows that The palatal unaspirated sound c is changed into an aspirated sound c^h sound at the syllable initial position without changing the meaning in the first entry and the bilabial plosive sound p is also changed into the aspirated p^h sound at the syllable initial and positions are shown in the table's respectively second and third-fourth entries.

The aspiration in the language affects the definition of the term or word. The existance

Table 3.9: Aspiration in Kishoreganj Dialect

Rules	Standard Bengali [IPA]	Kishoreganj Dialect [IPA]
/c/ - /cʰ/	চার [cɐr]	ছাইর [cʱɐɪr]
/p/ - /pʰ/	পঞ্চাশ [poncaʃ]	ফইঞ্চাশ [pʱo̯ɪncaʃ]
/p/ - /pʰ/	বাটপার [bɐtpɐr]	বাটফার [bɐtpʱar]
/p/ - /pʰ/	বাপ [bɐp]	বাফ [bapʱ]

Table 3.10: Aspiration in Narail Dialect

Rules	Standard Bengali [IPA]	Narail Dialect [IPA]
/p/ - /p^h/	পারে [pɐre]	ফারে [p^hɐre]
/p/ - /p^h/	প্রত্যেক [proṯṯek]	ফত্যেক [p^hoṯṯek]
/p/ - /p^h/	তারপরে [t̪ɐrpɔre]	তারফরে [t̪ɐrp^hɔre]
/c/ - /c^h/	খরচ [k^hɔroc]	খরছ [k^hɔroc^h]

of the same case is also observed in the dialect of Narail. However, the speaker of Narail also has a tendency to alter the unaspirated sound into an aspirated sound where this kind of change does not affect the meaning of the word. The examples are shown in the table 3.10. In the first two entries, the bilabial plosive sound p is changed into the aspirated p^h sound at the syllable initial position. In the third entry of the table, the bilabial plosive sound p is changed into the aspirated p^h sound at the syllable middle position and in the final entry, the palatal unaspirated sound c is changed into an aspirated sound c^h without changing the meaning.

Chittagonian consonant is different from the Standard Bengali consonant. Some examples can be seen in the table 3.11. In the first and second entry, the bilabial plosive sound p and k is changed into the aspirated p^h and $k^ɦ$ sound at the syllable initial position. The bilabial plosive sound p is changed into the aspirated p^h sound at the syllable medial position seen in the third entry of the same table and the bilabial plosive sound p is changed into the aspirated p^h sound at the syllable final position is seen in the last entry of the same table.

Table 3.11: Chittagong dialect aspiration example

Rules	Voiceless aspiration [IPA]	Meaning in English	Voiced aspiration [IPA]
/p/ - /p^h/	পরীক্ষা [$p^ɦ$orikka]	Exam	ফরিক্কা [porɪkk^{ɦ}a]
/k/ - /k^h/	কালা [kɐlɐ]	Black	খালা [$k^ɦ$ɐlɐ]
/p/ - /p^h/	তারপরে [t̪ɐrpɔre]	Then	তারফরে [t̪ɐrp^hɔre]
/p/ - /p^h/	মাপ [mɐp]	Measure	মাফ [mɐp^h]

One different characteristic of the Narsingdi dialect is the number of unaspirated consonants compared to aspirated ones. This indicates that, in contrast to certain other Bengali dialects, there is less of a noticeable burst of air (aspiration) when producing voiceless

stops like k, g, and c. Hence, in contrast to standard Bengali or other regional dialects, the speakers of the Narsingdi dialect pronounce words with a comparatively softer and less aspirated quality for these consonant sounds. Though Some $p^{ɦ}$ sounds are heard in the pronunciation of Narsingdi speakers. Some examples are in the table 3.12.
The bilabial plosive sound p is changed into the aspirated $p^{ɦ}$ sound at the syllable initial position in the first entry of the table. The same changes are made respectively at the syllable medial and final position in the second and third entries of the same table.

Table 3.12: Narsingdi dialect aspiration example

Rules	Voiceless aspiration [IPA]	Meaning in English	Voiced aspiration [IPA]
/p/ - /$p^{ɦ}$/	পড়া [pɔɽɐ]	Read	ফরা [$p^{ɦ}$ɔɽɐ]
/p/ - /$p^{ɦ}$/	কাপড় [kɐpoɽ]	Cloths	কাফড় [kɐ$p^{ɦ}$oɽ]
/p/ - /$p^{ɦ}$/	কাপ [kap]	Cup	কাফ [kɐ$p^{ɦ}$]

3.2.4 Gemination

Gemination refers to the prolonged pronunciation of a consonant due to its consecutive occurrence in a sentence [15]. For instance, বাদ দাও [bad̪ d̪a͡o̯] can also be pronounced as বাদ্দাও [bad̪d̪a͡o̯], with the latter being a geminated form. The distinction between these two pronunciations is subtle in spoken language but significant in written text. In Bengali, gemination can occur when similar phonemes are shared between words and within a single word, with a higher frequency between two consecutive words.
While this linguistic feature presents challenges for determining word boundaries in ASR algorithms for various languages [17], the specific modeling challenges it poses in Bengali have not been extensively explored in existing literature.
Gemination involves the extended pronunciation of a consonant sound in comparison to its typical length. The precise pronunciation and the use of germination may vary from dialect to dialect. We can see some examples of geminition for Rangpur, Kishoreganj, Narail, Chittagong and Narsingdi respectively in the table 3.13.

Table 3.13: Gemination examples of different regional dialects with IPA

Rules	Standard Bengali [IPA]	Dialect [IPA]
Rangpur		
/l + l/	আল্লাহ [ɐllɐh]	আল্লা [ɐllɐ]
/t̪ + t̪/	ওইতো [oɪt̪o]	ওইত্ত [oɪt̪̪o]
/t + t/	এখানে [ekʰɐne]	এট্টে [eɪtte]
/t̪ + t̪/	ওইগুলো [oɪgulɐ]	ওইল্লা [oɪllɐ]
/n + n/	শুনলাম [ʃunlɐm]	শুন্নু [ʃunnu]
Kishoreganj		
/n + n/	কিনে [kɪne]	কিন্না kɪnnɐ
/t̪ + t̪/	কেটে [kete]	কাইট্টা [kɐi̯̪tta]
/k + k/	অপেক্ষা [opekkʱɐ]	অপেক্কা [opekkɐ]
/c + c/	বেচে [bece]	বেইচ্চা [bei̯cca]
/t̪ + t̪/	মথ্যা [mɪt̪̪ʱɐ]	মিত্তা [mɪt̪̪ɐ]
Narail		
/m + m/	জন্মের [ɟɔnmeɾ]	জম্মের ɟɔmmeɾ
/t̪ + t̪/	করতি [koɾt̪ɪ]	কত্তি [kot̪̪ɪ]
/t + t/	একটা [ɛktɐ]	এট্টা [ɛttɐ]
/d̪ + d̪/	মধ্যে [mod̪̪ʱe]	মদ্দি [mod̪̪ɪ]
/n + n/	এমনে [emne]	এন্নে [enne]
Chittagong		
/ɟ + ɟ/	করেছি [koɾecʱɪ]	গইজ্জি goi̯̪ɟɟɪ
/l + l/	এখন [ɛkʱon]	এল্লা [ɛllɐ]
/d̪ + d̪/	মধ্যে [mod̪̪ʱe]	মইদ্দে [moi̯̪d̪̪de]
/k + k/	থেকে [tʱeke]	থাইক্কি [tʱɐi̯kkɪ]
/c + c/	আচ্ছা [ɐccʱɐ]	আইচ্চা [ɐi̯ccʱɐ]
Narsingdi		
/n + n/	কিনে [kɪne]	কিন্না [kɪnnɐ]
/k + k/	ঢুকে [dʱuke]	ডুইক্কা [dui̯kka]
/m + m/	এভাবে [ebʱɐbe]	এম্মে [ɛmme]
/l + l/	খুলে [kʱule]	খুইল্লা [kʱui̯lla]
/t + t/	হেঁটে [ɦ̃ete]	হাইট্ট [hɐi̯tta]

3.2.5 Diphthong

A diphthong refers to the combination of two vowels within a single syllable. The semi-vowel component of the diphthong can be found either at the beginning or the end of the syllable. Therefore, a diphthong is essentially a linguistic fusion of vowels and a glide sound.

The precise count of Bengali diphthongs has varied among different experts [32].

Sukumar Sen [8] noted that the Bengali language comprises two diphthongs, namely ঐ and ঔ. These combinations of two sounds are represented in written form but do not fit the conventional definition of diphthongs. In linguistic terms, they are referred to as digraphs [12].

On the contrary, Suniti Kumar Chatterji [2] claims that there exists 25 diphthongs in standard Bengali. In contrast, Md. Abdul Hai [4] asserts that there exists a total of 31 diphthongs, categorizing them into regular and irregular ones which are summing up to respectively 19 and 12. However, he also argues that there are only 18 diphthongs, as noted by Zinat Imtiaz Ali [12], who asserts that there are 17 diphthongs in Bengali. The government-approved IPA website acknowledges the regular 19 diphthongs, but they have used the diphthong /ui̯/ twice and exluded the /eo̯/ diphthong from considering.

A large number of loaned words from English, Arabic, Farsi, and others are used in Standard Bengali. These words are also spoken in the regional dialects as well. In English words with diphthongs, pronunciation is influenced by the presence of schwa/ə/ usually containing the neutral, unstressed vowel sound when appearing in unstressed syllables.

This leads to subtle variations in how diphthongs are articulated. For example, ‘power’- in the word, the diphthong /aʊ/ is followed by the schwa sound in the unstressed syllable. Or for the word ‘water’, the first syllable may be reduced to a schwa sound, especially if it’s unstressed. It might sound like ”wuh-ter.” However, when these words are adapted by the Bengali speaker they will be pronounced like /pa.o̯$^{\widehat{w}}$ar/ /o$\widehat{\text{o̯}}$.tɐr/.

Bengali speakers adopt English diphthongs that do not contain schwa and the pronunciation tends to align with the native English pronunciation. For example, হাই ’high’ is transcribed in the Bengali as /h$\widehat{\text{ɐi̯}}$/, boil as /b$\widehat{\text{ɔi̯}}$l/, and time as /t$\widehat{\text{ɐi̯}}$m/.

The table 3.14 shows how a standard Bengali diphthong deviates region-wise. The diphthong entry cells that are left blank mean that those diphthongs do not exist in my curated

dataset or perhaps even in that regional dialect.

Note: Abdul Hai [4] categorized the regular diphthong /eo̯/ for words such as কেও, পেও and the irregular diphthong /eo̯/ for words like চেয়ো, পেয়ো. These words have orthographic differences but rarely do they have distinctions in their pronunciation. Which are marked as * in the table 3.14

Table 3.14: Dipthong table of different regions The IPA transcription follows [89]

Serial	Standard Bengali	IPA	Rangpur	Kishoreganj	Narail	Chittagong	Narsingdi
Regular							
1.	ইই̯	ɪi̯	নিই [nɪi̯]	দিইয়া [d̪ɪi̯ɐ]	দিইয়া [d̪ɪi̯ɐ]	-	কাকিইয়ে [kɐi̯e]
2.	ইউ	ɪu̯	টিউবল [tɪu̯bɔl]	-	-	ইয়ুত [ɪu̯t]	-
3.	এই̯	ei̯	গেইলে [gei̯le]	এইডা [ei̯dɐ]	যেইডে [jei̯de]	এইক্কের [ei̯kkere]	এই [ei̯]
4.	এও*	eo̯*	-	এও [eo̯]	-	কেওরেত্তুন [keo̯rettun]	এও [eo̯]
5.	এউ̯	eu̯	-	কেউ [keu̯]	কেউ [keu̯]	কেউ [keu̯]	কেউ [keu̯]
6.	অ্যাও̯	ɛo̯	দেও [d̪ɛo̯]	দেও [d̪ɛo̯]	দেও [d̪ɛo̯]	-	দেও [d̪ɛo̯]
7.	অ্যায়̯	ɛe̯	-	দেয় [d̪ɛe̯]	দ্যায় [d̪ɛe̯]	-	দেয় [d̪ɛe̯]
8.	আই	ɐi̯	রাইত [rɐi̯t]	নাইগা [ɐi̯ga]	আইসে [ɐi̯je]	আইবো [ɐi̯bo]	আইলে [ɐi̯le]
9.	আএ	ɐe̯	এলায় [elɐe̯]	আয়ে [ɐe̯]	পায় [pɐe̯]	গায়ে [gɐe̯]	মায়ে [mɐe̯]
10.	আও̯	ɐo̯	ইয়াও [ɪɐo̯]	আও [ɐo̯]	আওনে [ɐo̯ne]	-	আওয়াজ [ɐo̯ʲɐɟ]
11.	আউ̯	ɐu̯	কাউরে [kɐu̯re]	আউজ্জা [ɐu̯ɟɟɐ]	আউলাত [ɐu̯lɐt]	আগাউ [ɐgɐu̯]	আউশ [ɐu̯ʃ]
12.	অয়̯	ɔe̯	কয় [kɔe̯]	অয় [ɔe̯]	অয় [ɔe̯]	নয় [nɔe̯]	অয় [ɔe̯]
13.	অও̯	ɔo̯	-	কও [kɔo̯]	কও [kɔo̯]	সওরি [sɔo̯ri]	অওয়া [ɔo̯ʲɐ]
14.	ওও	oo̯	ধোও [d̪ʰoo̯]	-	-	-	-
15.	উআ	ou̯	-	-	যৌবন [jou̯bɔn]	হউন [hou̯n]	বউ [bou̯]
16.	ওই̯	oi̯	কইলে [koi̯le]	মইদ্দে [moi̯d̪d̪e]	কইরে [koi̯re]	চৈত [coi̯t]	কইরা [koi̯rɐ]
17.	ওয়̯	oe̯	-	-	-	-	-
18.	উই̯	ui̯	থুইয়া [tʰui̯ɐ]	থুইতাম [tʰui̯tɐm]	দুইডে [d̪ui̯de]	উইদ্দো [ui̯d̪d̪o]	উইট্টা [ui̯ttɐ]
19.	উউ	uu̯	খুউব [kʰuu̯b]	রুউবাইন [ruu̯bɐi̯n]	-	-	খুউব [kʰuu̯b]
Irregular							
1.	অআ	ɔɐ	-	কয়া [kɔa͡]	-	মআল [mɔa͡l]	অয়াতে [ɔa͡te]
2.	ইআ	ɪɐ	বালিয়া [bɐlɪɐ͡]	বিয়া [bɪɐ͡]	দিয়া [d̪ɪɐ]	মিয়া [mɪɐ͡]	দিয়া [d̪ɪɐ͡]
3.	ইও	ɪo̯	-	অইয়ো [ɐɪo̯]	দিও [d̪ɪo̯]	কিওর [kɪo̯r]	দিওন [d̪ɪo̯]
4.	ইএ	ɪe̯	-	দিয়ে [d̪ɪe̯]	-	দিয়ে [d̪ɪe̯]	কিয়ের [kɪe̯r]
5.	এআ	eɐ̯	যেয়া [jeɐ̯]	নেয়া [neɐ̯]	দেয়ার [d̪eɐ̯r]	গেয়া [geɐ̯]	টেয়া [te͡]
6.	এয়ো*	eo*	-	-	-	-	-
7.	অ্যায়া	ɛɐ	-	ট্যায়া [tɛɐ͡]	দ্যায়াচ্ছে [d̪ɛɐ̯cc͡ʰ]	-	-
8.	ওআ	oɐ	-	হওয়া [hoɐ͡]	-	নোয়া [noɐ͡]	লওয়া [lɔoɐ͡]
9.	ওএ	oe̯	-	-	-	-	-
10.	উয়ে	ue̯	-	-	শুয়ে [ʃue̯]	ফুয়ে [pʰue̯]	-
11.	উআ	uɐ	-	দুয়া [d̪uɐ͡]	-	বুয়া [buɐ͡]	আঙুয়া [ɐggua͡]
12.	উও	uo̯	-	কুওন [kuo̯n]	-	উওর [uo̯r]	-

3.2.6 Triphthong

Triphthong is described as ”a combination of three vowel sounds, where the first vowel glides into the second, which, in turn, glides into the third,” as defined by

Barman et al. [20]. The English language contains triphthongs, which is a rare case in the Bengali language. According to Suniti Kumar Chatterji [1], there are triphthongs in the Bengali language such as aeoi, aie, aio, aei, aoi, aui, eie, eio, eao, eoi, euo, iei, ieo, iae, oei, ooi, oeo, oie, ogi, oeo, oai, oae, oui, seei, uie, uio, uei, ueo, uae, uao, uoe.

However, the triphthongs in the Bengali language do not exist in the same way they exist in English. In the case of English triphthongs, native Bengali speakers tends to convert it into a dipthong and completely avoid pronouncing the triphthong word. For example, in English, the word 'fire' is pronounced as /fʌɪə/, which in Bengali is transcribed as /fɐ͡j̯e̯.ɐr/. Cases like these are found in these words as well - 'hour' /aʊər/, which is pronounced as /ɐ.o͡w̯e̯r/, 'prayer' /preɪər/, pronounced as /pɾeʲ.ɐr/, 'pure' /pjʊr/ pronounced as /pɪ͡o̯r/. A Bengali speaker's pronunciation of words can vary based on specific contexts and regional accents. Even a standard native speaker may pronounce certain words differently depending on the situation, which could lead to variations in IPA transcription. Unless the transcription is based on audio data, ensuring accurate contextual transcription can be a challenge. There are instances in dialectal languages where two vowels can appear together in a word, but they are treated as separate syllables.

For example, the word থুইলেও [t̪ʰuɪleo] is used by the speaker of Rangpur. Here the vowel combination eo is part of the separate syllable as in t̪ʰuɪle.o. Another similar case can be seen in দিয়াও [d̪ɪɐo] which is used by the speaker of other regions as well. These cases are unlike true triphthongs.

Excluding these, many grammatical features exists in Bengali that make ASR and TTS modeling challenging in this language. Brahmic Schwa deletion ambiguity [45] i.e. without an appropriate language model, trailing vowel sounds dropped by many Bengali words are difficult to model. Related to this issue, modeling is more challenging for a great amount of homographs in Bengali [72] [41]. Although the markers of Nasality [6] is often explicitely included in the written forms but in practice, it is often skipped while pronouncing done by any native Bengali speakers. Also, it has been reported [7] that the phonological phrases in Bengali correspond to no plausible constituent of syntactic representation which results as a challenge when modeling speech ensuring proper prosody.

In Bengali speech, the phone(s) associated with glide, dipthongs and emphasizers often sounds similar. Where by definition, glides are sound that is phonetically similar to a vowel sound but operates as the syllable boundary. Dipthongs are the sounds formed when combining the two vowels and emphasizers or modifiers that enhances to give the

auxiliary emotional context.
As each of these have different written narratives, high contextual information is often required when differentiating them from each other and this also results as a challenge when modeling. [18].
An inevitable instance in Bengali sentence structure is the position swapping of morpheme in the sentence. Bengali Language follows the SOV pattern of sentence structure. But in the daily life discourse or in literature, this pattern often alters to SVO, VSO, and OVS. [35] [33]
Also, a great deal of foreign words from different languages are used by Bengali speakers that often has no standardized pronunciation and spelling. This results as a practical challenge when working with diversified real-world data [5]

3.3 Available SOTA ASR Models

Numerous models have been fine-tuned using Bengali corpora; however, there is a notable absence of datasets specifically focused on Bengali speech incorporating regional dialects. Consequently, none of these models have undergone fine-tuning in this context. Several of these models were employed to benchmark performance against our developed corpus. A detailed description of these models is provided in the following sections.

3.3.1 YellowKing

In July 2022, Bangladesh University of Engineering and Technology (BUET)'s Computer Science and Engineering (CSE) department initiated the first-ever Bengali Automatic Speech Recognition competition in Bangladesh, hosted on Kaggle under the title "DL Sprint" [1]. The competition spanned from July to September and aimed to accelerate research in AI models for Bengali speech recognition. The significance of this initiative lies in the vast audience and potential business and educational applications associated with this field. The competition utilized a corpus of recorded sentences provided by volunteers through the Mozilla Common Voice platform.

The competition was sponsored by the ICT Division of the Bangladesh Government, with a prize of $200,000$ Bangladeshi Taka awarded to the winning team.

[1] https://www.kaggle.com/competitions/dlsprint/overview

The competition consisted of two phases. In the first phase, participating teams used 400 hours of annotated Bengali speech data from Mozilla Common Voice to train their models. Teams competed to achieve optimal scores on a common voice test dataset. Upon completion of this phase, teams submitted their models and corresponding reports for evaluation. Performance was assessed on two distinct test sets, which remained confidential, accessible only to a select group of evaluators. The top-performing teams were subsequently invited to BUET CSE for the final phase, where they presented their work and received their awards. The judging panel, comprising industry professionals, government representatives, and academic experts, selected the winning model based on several criteria, including public and private Kaggle rankings, performance on a hidden test set, word error rate, runtime, and Kaggle score. The model named YellowKing[80] was ultimately chosen as the winner of the competition.
The project utilized a dataset comprising audio clips and corresponding transcripts, totaling $42,941$ data points. However, approximately 13% of these data points were deemed unreliable and subsequently excluded based on a voting mechanism. The audio files were standardized by converting them to a 16 kHz MP3 format, while transcript characters were transformed into integer hashes. The team employed the Transformers library, optimizing batch training and matched the length of the longest sequence by padding all the other sequences in a batch.

To enhance the efficiency of the training process, portions of the audio arrays with low amplitude values at the beginning and end were removed. The dataset was further refined by selecting only sound files with durations between 1 and 9 seconds, resulting in a final dataset of $36,919$ data points. The model was implemented using PyTorch, specifically leveraging a pre-trained Transformer model, *facebook/wav2vec 2.0-large-xlsr-53* fine-tuned for the task. The training procedure was conducted in two phases, ultimately achieving a Levenshtein Mean Distance Score of 2.60753.

The final model submission yielded a Connectionist Temporal Classification (CTC) training loss of 0.3172 and an evaluation WER of 0.2524. While further reductions in training loss were possible, the WER plateaued after approximately 77 epochs.

3.3.2 Google ASR

Google Automatic Speech Recognition (Google ASR) is an advanced speech recognition system developed by Google, serving as a crucial component of various Google services, including Google Assistant, Google Search, and Google Translate. This technology facilitates user interaction with devices through spoken language, transcribing speech into written text for a wide range of applications. Over time, Google ASR has undergone continuous enhancements, leading to increasingly seamless and accessible voice interactions for users globally. Its applications span multiple domains, such as voice assistants, search engines, transcription services, and voice-controlled devices. In essence, Google ASR is a powerful tool that significantly improves user interaction with Google's services and products via spoken language.

In our study, We utilized the Google ASR API to transcribe the test audio samples from our dataset dataset.

While Google ASR performed adequately on standard Bengali, its performance deteriorated significantly when processing regional Bengali dialects. For instance, the system exhibited high CER and WER when transcribing audio from districts like Chittagong (CER: 0.943, WER: 0.984) and Kishoreganj (CER: 0.893, WER: 0.948). The primary reason for this decline in performance is that Google ASR has been trained predominantly on standard Bengali audio. Consequently, when confronted with audio clips containing regional dialects and out-of-vocabulary (OOV) words unfamiliar to the model, its transcription accuracy was markedly reduced.

3.3.3 Wav2Vec 2.0 Large

XLS-R has demonstrated significant advancements over previous state-of-the-art models for a range of applications, such as speech recognition, translation, diarization, and even language identification as detailed in the official documentation **XLS-R** and note that, YellowKing[80] was a fine-tuned version of Wav2Vec 2.0 Large XLS-R.
For our study, We conducted inferences on the test data samples from our developed corpus using a fine-tuned version of Wav2Vec 2.0 Large XLS-R, as hosted by Auditi Das on Hugging Face [91]. The Common Voice 11.0 dataset [55] was utilized to fine-tune this model.
Among the various configurations and iterations tested, the version referred to as "Yel-

lowKing" emerged as the most stable and robust performer. This Wav2Vec 2.0 model was then evaluated using the entire test split to assess its performance.
Although the model performed adequately in standard Bengali, its performance declined when applied to the regional speech dataset across all districts. This decrease in accuracy can be attributed to the model's unfamiliarity with regional dialects, accents, and out-of-vocabulary (OOV) words.

3.3.4 Tugstugi (Whisper-Medium)

OpenAI's Whisper-medium [79] is a language model designed to support a wide range of languages, including Bengali. Whisper's capabilities extend to both transcription and translation of spoken content across multiple languages and dialects, thereby facilitating cross-linguistic communication and helping to overcome language barriers.
With training on a comprehensive dataset of $680,000$ hours of labeled speech data, Whisper-medium is well-equipped to handle a wide array of languages and dialects. This variant performs robustly across both monolingual and multilingual tasks, delivering strong results in various linguistic contexts and domains. Its flexibility enables it to convert spoken language into text and translate speech across languages, enhancing cross-lingual communication.

In this study, we conducted inference on test data samples from our developed corpus utilizing a fine-tuned version of Whisper Medium, designated as Tugstugi [88], which is hosted on Hugging Face. Tugstugi emerged as the winner of the Bengali.AI Speech Recognition Competition **bengaliai-speech**. This model was fine-tuned using two distinct speech corpora, including OOD-Speech: A Large Bengali Speech Recognition Dataset for Out-of-Distribution Benchmarking [87].

3.3.5 Hishab-Conformer

The Conformer architecture represents a deep learning framework specifically designed for ASR and NLP tasks. Conformer models have gained significant attention due to their effectiveness in processing sequential data and their ability to achieve superior performance across various ASR benchmarks. The Conformer framework incorporates several essential components, including convolutional layers, a self-attention mechanism, and supplementary elements derived from the Transformer architecture. This hybrid approach not only allows Conformer models to efficiently handle sequential data but also

Table 3.15: Fine-tuned Conformer Parameters

Parameters	Value
epochs	16
batch size	32
sampling rate	16kHz
use start end token	True
pin memory	True
number of workers	48
trim silence	False
max duration	18.5
min duration	0.2

ensures their adaptability to diverse linguistic and environmental conditions, making them a preferred choice for both academic research and industrial applications in ASR. Conformer-based ASR models have consistently demonstrated state-of-the-art performance across numerous benchmarks, highlighting their efficacy in comprehending and transcribing spoken language.

In our study, we benchmarked a fine-tuned FastConformer model [86] developed by Hishab. This model was trained on the nearly $20,000$-hour MegaBNSpeech corpus, created by the authors. The model, based on the NeMo Toolkit and utilizing a Conformer-CTC architecture, was further fine-tuned on $4,000$ hours of transcribed YouTube data. A byte-pair encoding tokenizer [59] was initially developed utilizing the train split's transcriptions.

The encoder's weight initialization were done utilizing the pre-trained weights of Nemo English ASR during training. The training parameters are shown in the table 3.15

Chapter 4

Methodology

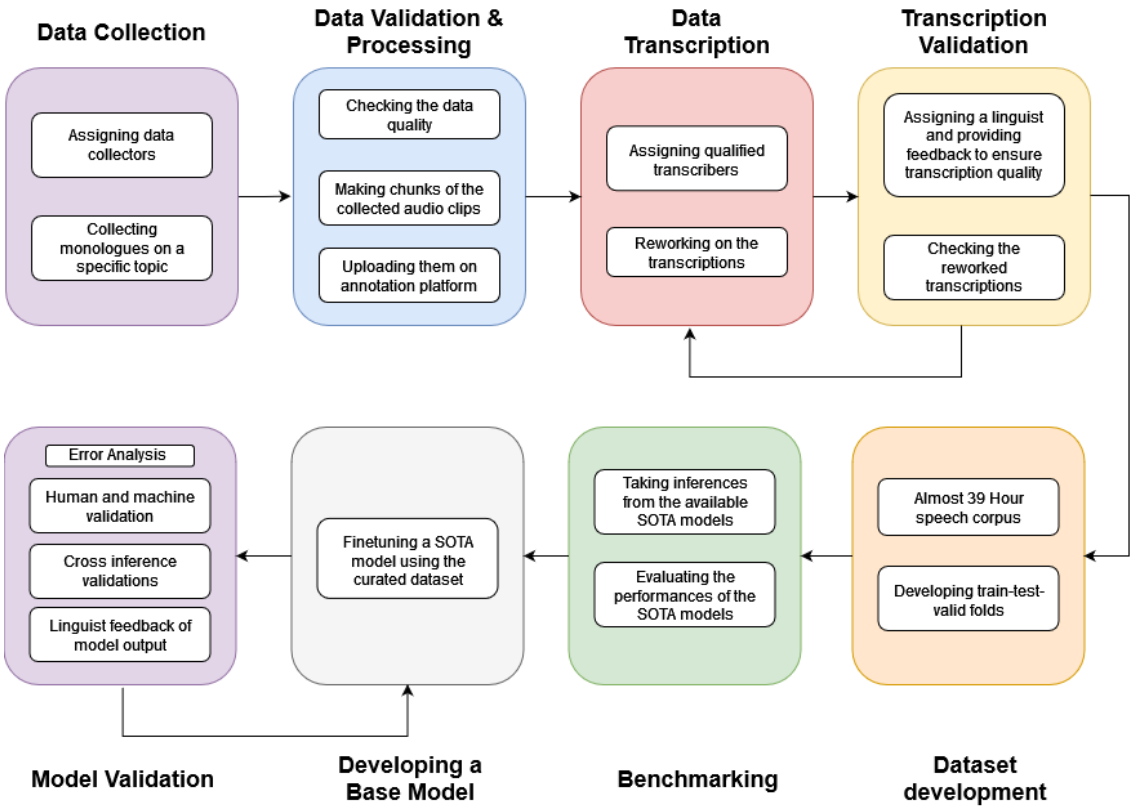

Figure 4.1: Top-level overview of the proposed methodology

In this research, we developed a speech corpus with a specific focus on regional dialects, along with a speech-to-text (STT) model designed to transcribe Bengali speech featuring regional dialectal variations, adhering to a standard orthography established by linguists. The pipeline depicted in Figure 4.1 delineates the comprehensive sequential steps to achieve our objective. This approach serves as an accelerated method for automating STT processes for low-resource languages, particularly those with numerous dialectal

variations and substantial linguistic diversity.

The diagram referenced above illustrates the comprehensive workflow for our project. In light of the absence of publicly available Bengali speech corpora specifically tailored to regional accents and dialects, we have undertaken the initiative to construct a corpus independently. This process involved the meticulous collection of data following established protocols designed to encourage spontaneous speech. Additionally, A compilation of monologues was obtained from individuals who were prompted to express their thoughts and sentiments without any restrictions or diffidence.

Following the audio data validation and processing phase, the recruitment of proficient transcribers from the different geographical regions relevant to the data was executed meticulously. This deliberate choice aimed to safeguard and enhance the quality of transcription by leveraging the transcribers' intimate familiarity with the nuances of accent and dialect present in the speech data. To ensure their linguistic competence, a comprehensive transcription examination was administered, encompassing both standard Bengali and the respective regional variations.

To further strengthen the quality assurance process, a dedicated linguist was appointed to rigorously review the transcriptions and provide constructive feedback. The transcribers then incorporated the suggested revisions and refinements, thereby fortifying the overall quality of the transcribed materials.

Following the aforementioned processes, we curated the dataset and employed an 80:10:10 partitioning strategy to divide it into training, testing, and validation subsets.

To optimize our model's performance, we fine-tuned a state-of-the-art (SOTA) model using our proprietary dataset as the training corpus. To evaluate the efficacy and competitiveness of our model, we conducted benchmark assessments by inputting our dataset into various other SOTA models and subsequently evaluating their transcription capabilities.

Additionally, we meticulously engineered an accent classifier capable of identifying and categorizing speech data based on the presence of regional accents and dialects.

Currently, our dataset comprises nearly 39 hours of speech data from five distinct districts within Bangladesh: Rangpur, Kishoreganj, Narail, Chittagong, and Narsingdi.

To summarize, through this work, we accomplished the following:

- Compiled nearly 39 hours of speech corpora specifically tailored to regional accents and dialects.
- Conducted both human and machine validation of the dataset.
- Generated benchmark results from various publicly available ASR models.

4.1 Data collection and validation

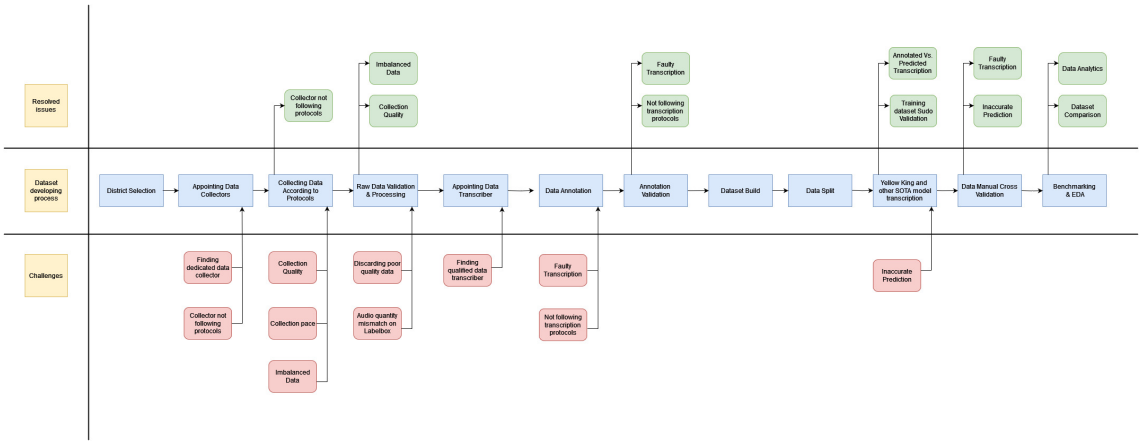

Figure 4.2: Detailed Workflow Diagram of Data Collection

Our objective was to develop a speech corpus that would enable any ASR model trained on it to accurately recognize Bengali speech from a variety of speakers across different topics, without exhibiting bias. To achieve this goal, several features were incorporated to ensure diversification, as outlined below.

1. Voice Diversification: To encompass a diverse array of voices and prevent bias towards any particular voice, we aimed to gather data from a wide range of individuals. Each audio clip was maintained at approximately ten minutes in duration. We targeted collecting at least 5 hours of data from each district.
2. Gender Diversification: To achieve a balanced gender representation, we aimed for a 50:50 ratio in the voice dataset. For a dataset comprising 5 hours of data, involving 30 individuals with each contributing 10 minutes of speech, we included 15 distinct male voices and 15 distinct female voices.
3. Age Diversification: Considering the changes in a person's voice with age, primarily due to reduced lung capacity and declining muscle strength and tone [90],

we incorporated age-related diversification into our speech corpus. This ensures that the ASR system trained on this corpus can effectively recognize the speech of speakers across various age groups.

4. Topic Diversification: To train an ASR system capable of recognizing speech across a wide range of topics and vocabulary, data collectors were guided to engage in conversations on everyday subjects. They were provided with a comprehensive list of topics, including sports, education, politics, family life, economics, and more. This approach aimed to ensure the system's proficiency in recognizing speech on virtually any topic.

5. Geographical Diversification: Geography significantly impacts accent adjustment [92], affecting various aspects of human interaction and linguistic diversity.

Incorporating these diversification features, we initiated data collection from various regions. The procedure is illustrated in Figure 4.2, with details discussed below.

4.1.1 District Selection

Districts were selected based on two criteria:

1. Available Acquaintances: Districts where personal connections existed were chosen to receive direct assistance from known individuals or to delegate tasks as needed.

2. Geographical Distance: To ensure geographic diversity within the same language, districts or regions located at considerable distances from one another were selected, as languages tend to remain relatively consistent over short distances.

4.1.2 Appointing Data Collectors

In-person visits were avoided to obtain spontaneous speech samples. Additionally, being non-native to the region, there was a concern that speakers might not use their native accent while conversing with us and might not be as open compared to interactions with someone native to the area.

4.1.2.1 Challenges

1. Finding Dedicated Data Collectors: Identifying a dedicated data collector who adhered to all collection protocols was crucial to ensure the five types of diversification outlined above.

2. Collector Adherence to Protocols: Ensuring that the data collector diligently followed the protocols was essential for achieving the desired diversification in the speech corpus. Deviations in data collection could result in ASR models trained on this corpus not meeting the expected performance standards.

4.1.3 Collecting Data According to Protocols

4.1.3.1 Resolutions

1. Collector Adherence to Protocols: Through manual validation of data collection, we could identify and correct potential data imbalances. This allowed us to guide the designated data collector in acquiring new data to ensure a balanced dataset.

4.1.3.2 Challenges

1. Collection Quality: The use of phone recorders by the designated data collector posed challenges such as low volume due to the device's distance from the speaker's mouth and the presence of background noise in the recordings.

2. Collection Pace: The rate of data collection could vary daily due to factors like weather conditions, illness, and personal commitments, affecting the consistency of data acquisition.

3. Imbalanced Data: Without adequate supervision during data collection, there was a risk of the dataset becoming imbalanced in various ways.

4.1.4 Raw Data Validation and Processing

4.1.4.1 Resolutions

1. Collection Quality: At this stage, we reviewed the audio clips, categorizing those with consistently low speaker volumes and identifying clips with significant background noise.

2. Imbalanced Data: During manual validation, we precisely identified the nature of data imbalances and devised methods to rectify them.

4.1.4.2 Challenges

1. Discarding Poor Quality Data: Upon thorough examination of all collected data, we excluded recordings that did not adhere to the specified protocols or exhibited notably poor quality.

2. Audio Quantity Mismatch on Labelbox: Following the segmentation of audio using the VAD algorithm, we force-split any segments exceeding 30 seconds, ensuring that all audio chunks conformed to this duration limit. We then processed all the samples through a filter designed to eliminate two specific types of chunks: those with sound activity but no speech, and those with speech that is inaudible. Subsequently, we uploaded these samples to the data annotation platform, Labelbox [94], in a single batch. Occasionally, a bug may prevent all chunks from being uploaded simultaneously without providing notifications. To address this, we had to run a script to verify that all chunks were successfully uploaded to Labelbox.

4.1.5 Appointing Data Transcribers

4.1.5.1 Challenges

1. Finding Dedicated Data Transcribers: We hired transcribers who were residents of the same district where the data was collected. This choice provided them with an advantage as their ears were accustomed to the local accents and dialects. Applicants for the data transcription position underwent assessments with the assistance of a linguist who evaluated their scripts. The linguist selected candidates who demonstrated a strong command of Bengali spelling, grammar, and proficiency in their native accent. Following a thorough evaluation process, We selected several transcribers from the region to transcribe the audio data.

4.1.6 Data Annotation

I used the data annotation platform Labelbox to label the data. Transcribers were granted access to the data along with specific instructions and protocols provided by the linguist on handling the spellings of any out-of-vocabulary words. A screenshot of the interface of this annotation platform is shown in Figure 4.3.

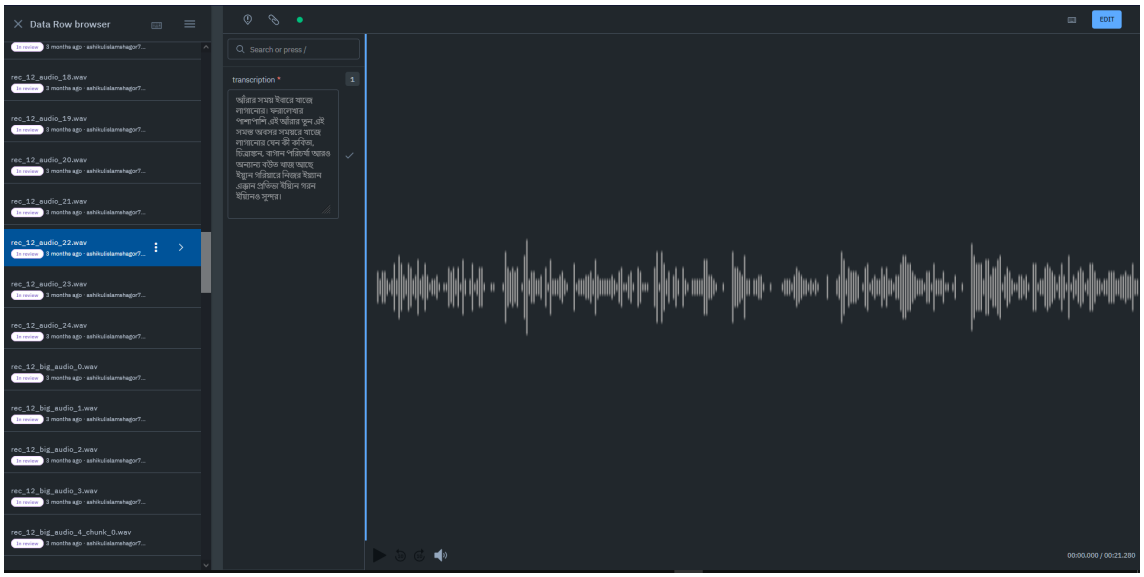

Figure 4.3: Interface of the annotation platform Labelbox

4.1.6.1 Challenges

1. Faulty Transcription: Despite clear instructions and adherence to established protocols, human annotators may still make errors during the transcription process, even if they are minor and unintentional.

2. Not Following Transcription Protocols: Certain individuals might become confused and fail to grasp the protocols, resulting in errors during the transcription process.

4.1.7 Annotation Validation

During the annotation process, a linguist diligently verifies spellings and transcriptions. Nonetheless, despite these efforts, occasional errors or inaccurate transcriptions may persist.

4.1.7.1 Resolutions

1. Faulty Transcription: Once all transcriptions for a specific region are complete, they are consolidated into a CSV file along with relevant information. This file is then shared with linguists for a review and feedback process. Subsequently, transcribers revisit and rectify any erroneous transcriptions based on the linguist's feedback. This final review ensures that the dataset is accurate before its completion.

2. Not Following Transcription Protocols: If a transcriber encounters confusion or makes errors, they will rectify them through discussion with the linguist or by implementing the linguist's feedback.

4.1.8 Dataset Build

After completing all transcriptions of the audio files, we removed entries that were null or had other problematic issues. This process resulted in a dataset that includes all relevant information for each audio clip, along with their transcriptions.

4.1.9 Dataset Split

After building the dataset, We performed 80:10;10 train-test-valid splits of the entire dataset.

4.1.10 SOTA Model transcription inferences

In this section, we performed inference on all audio samples within our dataset using cutting-edge models such as Google ASR, Wav2Vec2 large, A fine-tuned FastConformer which the authors claimed to have trained using 20,000 hours of pseudo-labeled Bengali speech data named, YellowKing (Kaggle Competition winner on the Bengali Common Voice Speech Dataset), and Tugstugi (Kaggle Competition winner on the Out-of-Domain Speech dataset). Additionally, we developed a base model by fine-tuning a pre-trained model using our proprietary training dataset.

4.1.10.1 Resolutions

1. Annotated Vs. Predicted Transcription: We computed the Word Error Rate (WER) and Character Error Rate (CER) scores for each of these inferences and our own fine-tuned model by comparing them with the human annotations in the dataset.

2. Training dataset sudo validation: During manual validation conducted by a linguist, we identified data imbalances and implemented corrective measures.

4.1.10.2 Challenges

1. Inaccurate Predictions:

We assigned a linguist to cross-check and validate the model's predictions based on several criteria:

- Diffs.: Comparison of the model's predictions with the human annotator's transcription to assess substantial discrepancies.
- Word Error Rate (WER): Quantifies the percentage of incorrect words in the predicted text compared to the human transcript.
- Character Error Rate (CER): Measures the percentage of inaccurate characters in the predicted text.
- Word Insertion: Number of additional words not present in the human transcription.
- Word Deletion: Number of words missing from the human transcription.
- Word Insertion and Deletion Total: Combined count of inserted and deleted words by the models.

Based on these criteria, samples were categorized during the manual validation process into

- Incomplete Sentence: Instances where the ASR models failed to transcribe all spoken words in the audio clip.
- Incorrect Sentence: Cases where the ASR models transcribed incorrect words or included spelling mistakes in their transcriptions.

4.1.11 Data Manual Cross Validation:

4.1.11.1 Resolutions

1. Faulty Transcription: At this stage, the occurrence of faulty transcriptions is minimal. Any remaining inaccuracies are identified during the comparison of model predictions with ground truths, and the annotators promptly resolve them.
2. Inaccurate predictions: Upon reviewing the model predictions, identified issues are documented and addressed through appropriate corrective actions. After resolving these issues, the models are retrained to improve prediction accuracy.

4.1.12 Benchmarking and EDA

After obtaining all predictions, we proceeded to benchmark them against predictions made by other state-of-the-art (SOTA) models developed in previous steps.

4.1.12.1 Resolutions

1. Data Analysis: In this phase, comprehensive data analysis was conducted to assess metrics such as total correctly predicted words and characters. Additionally, visualizations were generated to provide insights into the dataset's performance.

2. Dataset Comparison: We compared our dataset with other available datasets by training two different models under identical parameters on each dataset. Subsequently, we evaluated and compared their predictions to determine the suitability of our dataset for training models specific to this task.

For the modeling process, the dataset was divided and pre-processed before being input into the model. The specifics of these steps are detailed below.

4.2 Dataset Split

The entire dataset, totaling almost 39 hours, was divided into train, test, and validation splits with a ratio of 80:10:10. The dataset consisted of $6,667$, 838, and 828 samples, totaling approximately 31:03:55, 3:51:11, and 3:51:43 hours of data, respectively. During this division, care was taken to ensure that recordings from a single speaker were included in only one of the splits. This approach facilitates a fair assessment of speaker generalization.

4.3 Modeling

We fine-tuned a large Wav2vec2.0 model [16], referred to as our proposed model, PX5. PX5 was fine-tuned from scratch for 100 epochs, with EarlyStoppingCallback enabled, and training concluded after 31 epochs.

4.3.1 Proposed Methodology

The standard implementation of Wav2vec2.0, in conjunction with the CTC algorithm, does not require a language model head or dictionary for transcription decoding. Models

like Wav2vec2.0 utilize the Wav2Vec2Tokenizer, which typically performs tokenization at the subword level as shown in the simplified architecture in Fig 4.4.

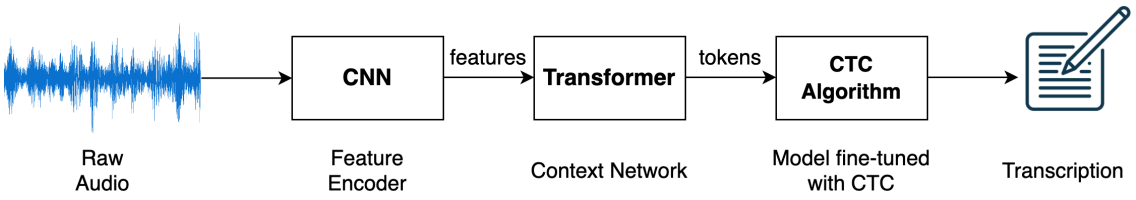

Figure 4.4: Simplified architecture of Wav2vec 2.0

While subword units can represent phonemes, syllables, or smaller linguistic elements, they effectively capture speech patterns and enhance the overall performance of the model in speech recognition/processing tasks. However, due to the significant diversity in local dialects and non-standard vocabularies in regional speech, we introduced a character-level n-gram language model. The architecture of our proposed methodology is illustrated in Fig 4.5.

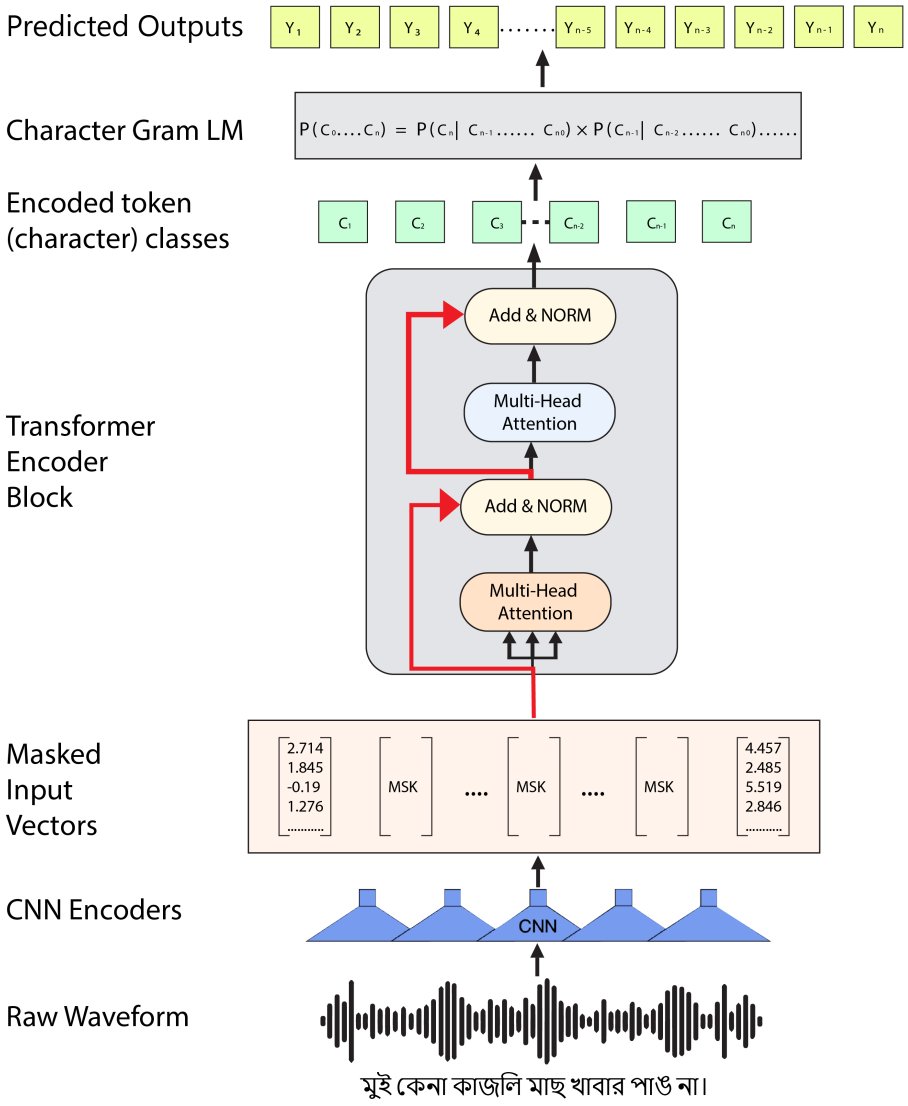

Figure 4.5: PX5 model architecture

This ASR system leverages self-supervised learning and transformers to efficiently decode audio signals into textual characters, making it particularly well-suited for low-resource languages like Bengali. The CNN encoders focus on feature extraction from the raw audio signal, while the masked modeling and transformer-based context capturing enable robust representations. The character gram language model further refines the output by modeling dependencies between characters, ensuring coherent transcriptions aligned with Bengali phonetics and grammar. A detailed breakdown of each component of this system is described below.

1. Raw Waveform (Input): The system begins by receiving the raw audio waveform as input, representing the speech signal of the speaker. The waveform is a continuous-time signal containing rich acoustic information, which must be pro-

cessed into discrete feature vectors for further analysis.

2. CNN Encoders (Feature Extraction): The raw waveform is passed through a series of CNN layers, which act as feature extractors. During this phase, the model learns a latent representation of audio, encoding it into high-level features that are abstracted from the raw signal converted by the CNN encoders. These features capture important information from the input audio, like speech patterns, and reduce the temporal dimension of the input, enabling the extraction of essential features such as phonetic information and temporal dependencies and are passed to the fine-tuning phase. However, they do not explicitly map to acoustic units like phonemes or syllables.

   These feature vectors form the backbone of the speech encoding process, feeding into the subsequent layers for more abstract representation.

3. Masked Input Vectors (Self-Supervised Learning): In this stage, portions of the feature vectors are masked (hidden), indicated by the ”MSK” markers. This is an essential part of the self-supervised pretraining phase in models like wav2vec 2.0. The model is tasked with predicting the masked portions of the input, forcing it to learn robust and informative representations of the speech signal without relying on labeled data. This method allows for efficient learning from large amounts of unannotated speech data, which is particularly beneficial for low-resource languages like Bengali.

4. Transformer Encoder Block (Contextual Representation): The masked input vectors are fed into a Transformer Encoder Block. The transformer architecture, featuring Multi-Head Attention mechanisms, enables the model to capture long-range dependencies within the speech sequence. It effectively models the global context by attending to different parts of the input sequence simultaneously.

   This is crucial in ASR systems as speech sounds can have contextual dependencies that are far apart in time. By employing multiple attention heads, the transformer encoder can focus on various parts of the input sequence and integrate contextual information from diverse perspectives.

   The Add and Normalize layers ensure that the model retains stable gradients during training and maintain layer consistency after the multi-head attention operations.

5. Encoded Token (Character Classes): The output from the transformer encoder block is a sequence of context-rich encoded tokens. These tokens represent high-

level abstractions of the input speech signal and are mapped to a specific class, which in this case, corresponds to characters in the Bengali alphabet. The model tries to predict the next character based on the acoustic features, and decoding happens on a per-character basis, where each encoded token translates to a character from the transcript. Each token is representing learned latent speech representations, an embedding containing both local acoustic information and global context, suitable for decoding into text.

6. Character Gram Language Model (LM): A character-level n-gram language model is used to decode the sequence of encoded tokens into a sequence of characters. This model estimates the probability of a character sequence given the context. The equation shown represents the joint probability of the character sequence $C_1, C_2, \ldots, C_n$ being the correct transcription, which is computed as the product of individual conditional probabilities:

$$\begin{aligned} P(C_1, C_2, \ldots, C_n) = P(C_1) \times P(C_2 \mid C_1) \times P(C_3 \mid C_1, C_2) \times \ldots \\ \times P(C_n \mid C_1, C_2, \ldots, C_{n-1}) \end{aligned} \tag{4.1}$$

   This ensures that the decoded output respects the character-level dependencies typical in Bengali regional scripts, improving the quality of transcription.

7. Predicted Outputs (Textual Representation): Finally, the system outputs the predicted sequence of characters, which forms the textual representation of the input speech. The model directly maps from audio features to characters as outputs without incorporating any acoustic unit discovery system. In this case, the predicted characters represent Bengali regional dialect texts, decoding the original audio input into a readable transcription in the target language.

4.4 Benchmarking

For benchmarking our fine-tuned model, We picked different state-of-the-art ASR models for Bengali. We involved popular speech API, and existing benchmark models used by the community. We do diverse benchmarking on our curated dataset. As the popular speech API, we used the Conformer-based $\mathbf{conformerModel}$ Google's speech-to-text cloud service API.

Google's cloud speech-to-text API was employed in its default configuration, without any fine-tuning. In contrast, a fine-tuned model [91] of Meta's Wav2Vec 2.0, trained using the Bengali Common Voice Speech Dataset [55], was used for benchmarking purposes. We also benchmarked two competition winner models: Yellowking [80] (The winner of the DLSprint Competition **dlsprint**) and Tugstugi [88] (The winner of Bengali.AI Speech Recognition Competition **bengaliai-speech**). Yellowking is a Wav2vec2 large [56] based and Tugstugi is a Whisper Medium [79] based finetuned model which was fine-tuned on 2 different speech corpus. Yellowking was fine-tuned on Bengali Common Voices Dataset [75] for about 70 hours on Kaggle GPUs and Tugstugi was fine-tuned on OOD-Speech: A Large Bengali Speech Recognition Dataset for Out-of-Distribution Benchmarking. [87] speech corpus. We also evaluated the Hishab Conformer [86] model which the authors have claimed to have trained on 20k hours of Pseudolabelled Bengali speech data.

We collected the inference results of all the available and traditional state-of-the-art deep learning models for automatic speech recognition using the samples from the test split of our curated dataset.

We evaluated their performance by calculating the average Word Error Rate (WER) and Character Error Rate (CER) of all the inference outputs from each of the models against human-annotated ground truth from regional sources and then compared it with our model performance. The outcomes are depicted in Table 6.2.

Chapter 5

Dataset Statistics, EDA and Feature Extractions

In this chapter, we delved deep into our developed speech corpus with regional Bengali dialects by showing all the statistics, exploratory data analysis, and relevant feature extraction from our corpus.

5.1 Dataset

This section shows all the corpus statistics of our corpus. In this section, we also illustrated how this regional speech corpus deviates from any standard Bengali speech corpus. We have used OOD-Speech: A Large Bengali Speech Recognition Dataset for Out-of-Distribution Benchmarking [87] as the reference corpus for standard Bengali.

5.1.1 Bengali Speech Corpus with regional dialects

The dataset encompasses five districts in Bangladesh: Rangpur, Kishoreganj, Narail, Chittagong, and Narsingdi. Predominantly comprising spontaneous speech, it also includes a small amount of monologues and phone-recorded conversations from each district. Further details specific to each district are provided in the follwing table 5.1.

Table 5.1: Overview of Bengali Speech Corpus Regional Dialects

Districts	Duration	Total chunks	Average chunk size	Type
Rangpur	6 Hour 1 Minutes	1,298	16.55 seconds	Spontaneous
Kishoreganj	9 Hour 36 Minutes	2,048	16.71 seconds	Spontaneous
Narail	8 Hour 36 Minutes	1,859	16.58 seconds	Spontaneous
Chittagong	8 Hour 11 Minutes	1,757	16.67 seconds	Spontaneous
Narsingdi	6 Hour 20 Minutes	1,371	16.51 seconds	Spontaneous
Total	38 Hours 46 Minutes	8,333	16.604 seconds	Spontaneous

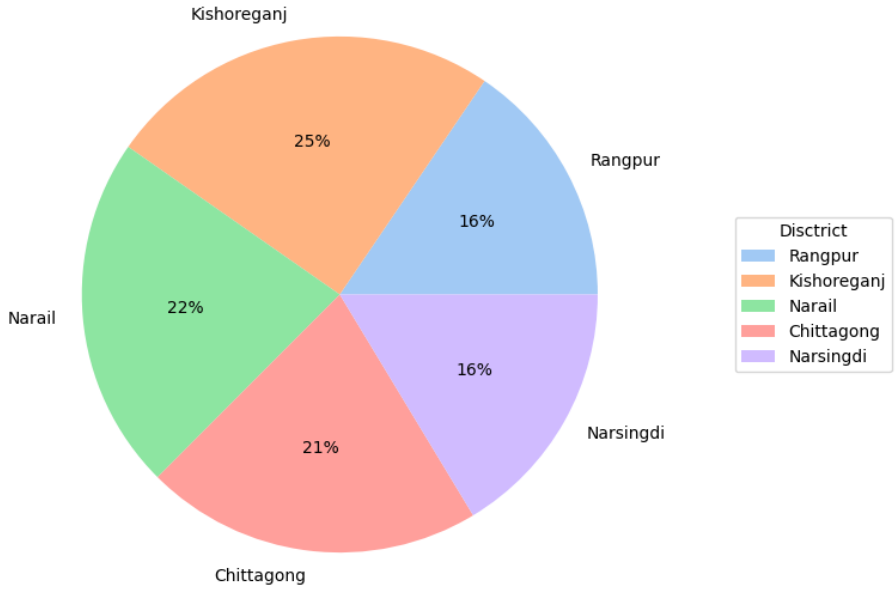

Figure 5.1: Pie chart of the corpus covering each district

5.1.2 About the corpus

The corpus has 8,333 chunk samples resulting from 178 speech recordings where the data was split into 80:10:10 train-test-valid split resulting in 31 hour 3 minutes, 3 hour 51 minutes, and 3 hour 51 minutes of audio data in each fold. In the metadata, Each chunk is accompanied by an 'External_ID', 'Districts', and 'Contents' which is the chunk transcription done by human annotators.

A detailed overview of the corpus is shown in the table 5.2

Table 5.2: Regional Speech Corpus Statistics.
↪ denotes subsets | WPM = Avg. Words Per Minute | WPS = Avg. Words Per Sample | H:M:S = Hour(s) : Minute(s) : Second(s)
| OOV = Words Out of Canonical Standard Bengali Vocabulary in comparison to the unique words of the corpus
| Annotation Complexity is measured by the time needed to annotate every unit of data.

Subset	**Samples**	**Duration (H:M:S)**	**Avg Rec. len.**	**WPM**	**WPS**	**Uniq. Words**	**OOV %**	**Characteristic Phones Pair Count**	**Avg. Phone length perc(%)**	**Annot. Compl.**
Bengali Regional Speech Dataset										
Die-alect Dataset (Cumulative)	8,333	38:46:49	16.754	132.429	36.965	41,878	70.125	12	0.982	118.5
↪Die-alect Train	6,667	31:03:55	16.774	133.68	37.355	36,177	68.834	–	0.987	54.6
↪Die-alect Test	838	3:51:11	16.553	130.61	36.004	8,244	55.471	–	0.999	221.02
↪Die-alect Valid	828	3:51:43	16.791	124.74	34.801	7,393	55.079	–	0.999	221.02
Region-wise breakdown of the dataset										
Rangpur (Cumulative)	1,298	6:00:57	16.685	134.022	37.171	8,827	51.263	5	0.989	21.286
↪ Rangpur Train	1,037	4:48:58	16.719	134.34	37.310	7,512	49.255	–	1.025	21.627
↪ Rangpur Test	130	0:35:15	16.269	122.657	33.023	1,642	38.733	–	1.038	19.901
↪ Rangpur Valid	131	0:36:44	16.822	146.25	40.191	1,761	38.785	–	1.038	19.901
Kishoreganj (Cumulative)	2,048	9:36:51	16.90	116.651	32.808	13,149	64.864	2	0.987	18.895
↪ Kishoreganj Train	1,638	7:42:30	16.941	117.665	33.187	11,477	64.364	–	0.987	19.303
↪ Kishoreganj Test	206	0:58:58	16.55	112.835	31.129	3,469	50.403	–	0.998	17.245
↪ Kishoreganj Valid	204	0:55:23	16.287	108.327	29.206	1,778	53.993	–	0.998	17.245
Narail (Cumulative)	1,859	8:36:52	16.68	136.12	37.783	10,960	57.409	6	0.971	8.74
↪ Narail Train	1,488	6:52:29	16.63	138.442	38.332	9,478	56.552	–	0.971	8.576
↪ Narail Test	188	0:51:57	16.69	127.262	25.402	3,387	44.133	–	0.983	9.387
↪ Narail Valid	183	0:52:25	17.188	117.038	33.257	1,977	42.792	–	0.983	9.387
Chittagong (Cumulative)	1,757	8:11:47	16.794	134.189	37.50	11,105	63.503	6	0.980	13.35
↪ Chittagong Train	1,406	6:35:54	16.89	136.618	38.381	9,517	61.868	–	0.981	12.02
↪ Chittagong Test	177	0:47:26	16.26	125.010	33.901	3,379	63.482	–	0.992	18.91
↪ Chittagong Valid	174	0:48:27	16.707	115.583	31.885	1,915	56.397	–	0.992	18.91
Narsingdi (Cumulative)	1,371	6:20:23	16.65	148.597	41.187	10,745	53.160	2	0.982	29.228
↪ Narsingdi Train	1,098	5:04:04	16.616	148.0	40.976	9,160	52.336	–	0.982	28.319
↪ Narsingdi Test	137	0:37:35	16.59	150.338	41.560	3,265	37.579	–	0.992	32.843
↪ Narsingdi Valid	136	0:38:44	17.091	156.789	43.809	1,718	39.464	–	0.992	32.843

The phoneme deviation of the regional Bengali from the standard Bengali can be observed in the table 5.3 in (Standard Bengali, Regional Bengali) template.

Table 5.3: Regional Phoneme Characteristics Pairs (Standard, Region)

Subset	Phoneme characteristic pairs	Count
Rangpur	(h, ^{h}), (h, ^{fi}), (k, t), (t, n), (n, p)	5
Kishoreganj	(^{h}, h), (ɟ, z)	2
Narail	(h, ^{h}), (h, ^{fi}), (k, t), (t, n), (n, p), (ɟ, z)	6
Chittagong	(c, ç), (cc, çç), (ɟ, z), (p, f), (p^{h}, f), (p^{h}, ff)	6
Narshingdi	(^{h}, h), (ɟ, z)	2
Die-alect Dataset	(h, ^{h}), (h,^{fi}), (k, t), (t, n), (n, p), (ɟ, z), (^{h}, h), (c, ç), (cc, çç), (p, f), (p^{h}, f), (p^{h}, ff)	12

5.1.3 Corpus statistics

The complete corpus consists of $8,333$ audio segments, totaling 38 hours, 46 minutes, and 49 seconds of data collected from five regions. The average segment length is 16.754 seconds, with a speech rate of 132.429 words per minute and an average of 36.965 words per segment. The corpus contains $41,878$ unique words, with 70.125% out of words vocabulary.

5.1.3.1 Train fold statistics

The train split comprises $6,667$ audio segments, amounting to 31 hours, 3 minutes, and 55 seconds of data collected from five regions. The average segment duration is 16.774 seconds, with a speech rate of 133.68 words per minute and an average of 37.355 words per segment. The corpus includes $36,177$ unique words, with 68.834% out of vocabulary words.

5.1.3.2 Test fold statistics

The test split consists of 838 audio segments, totaling 3 hours, 51 minutes, and 11 seconds of data gathered from five regions. The average segment duration is 16.553 seconds, with a speech rate of 130.61 words per minute and an average of 36.004 words per segment. The corpus contains $8,224$ unique words, with 55.471% out of vocabulary words.

5.1.3.3 Valid fold statistics

The validation split comprises 828 audio segments, totaling 3 hours, 51 minutes, and 43 seconds of data collected from five regions. The average segment duration is 16.791 seconds, with a speech rate of 124.74 words per minute and an average of 34.801 words per segment. The corpus includes $7,393$ unique words, with 55.079% out of vocabulary words.

5.1.4 Corpus Diversifications

5.1.5 Word and Grapheme Diversity

Figure 5.2 displays the different regions from which the data were collected, and table 5.3 illustrates the unique phonetic elements from each of those regions. In each phoneme pair, the first phoneme represents the conventional standard, while the second is the local variant used by inhabitants as a substitute for the standard phonetic sound.

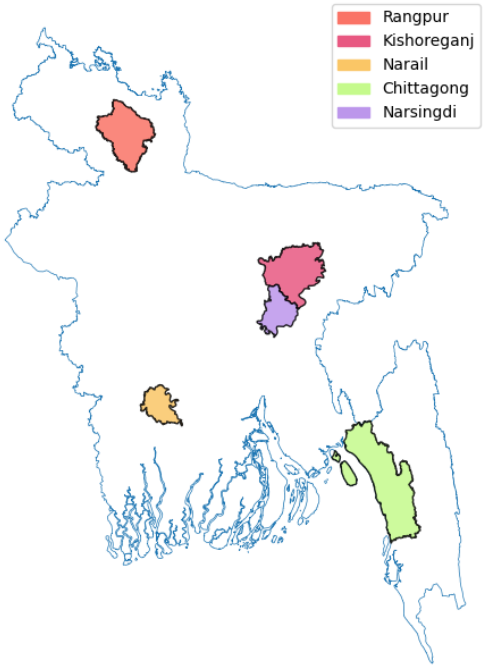

Figure 5.2: Mapped regions of the dialects along with reference point

Linguistic diversification is more conspicuous when two regions are geographically far apart from one other such as Rangpur, Chittagong, and Narail, which are farther from each other. Conversely, regions closer to each other or geographically adjacent regions such as Kishoreganj and Narsingdi, tend to exhibit similar levels of graphemic diversity. This suggests minimal linguistic variation within these neighboring areas, supporting the notion of mutual intelligibility [53].

Notably, the graphemic diversity observed in the Chittagong subregion is unique to that

area, reflecting a distinct collection of dialects specific to this region.
For a more comprehensive understanding, Table 5.4 demonstrates how an identical sentence is articulated differently across various regions, each employing a distinct set of words.

Table 5.4: Different pronunciation of the same sentence with IPA table

Subset	Bengali text	IPA Transcription
Standard Bengali	সামনে ইদ, ইদের কেনাকাটা করতে হবে ভাই!	ʃɐmne ɪd̪, ɪd̪er kenɐkatɐ kort̪e hɔbe bʱɐ͡ɪ!
Rangpur	সামনোত ইদ, ইদের কেনাকাটি কইরবার নাগবে, ভাই!	ʃɐmnot̪o ɪd̪, ɪd̪er kenɐkatɪ ko͡ɪ̯rbar nɐgbe, bʱɐ͡ɪ!
Kishoreganj	সামনে ঈদ, ঈদের কিনাকাডা করন লাগবো ভাই।	ʃɐmne ɪd̪, ɪd̪er kɪnɐkɐdɐ kɔron lɐgbo bʱɐ͡ɪ!
Narail	ছামনে ঈদ, ঈদির কিনাকাটা ওত্তি অবে ভাই!	cʰɐmne ɪd̪, ɪd̪ɪr kɪnɐkɐtɐ o̪tt̪ɪ ɔbe bʱɐ͡ɪ!
Chittagong	সামনো দি ঈদ, ঈদর কিনাকাটা গরন ফরিবু বাই।	ʃɐmno d̪ɪ ɪd̪, ɪd̪or kɪnɐkɐtɐ gɔron pʰorɪbu bɐ͡i̯!
Narsingdi	সামনে ইদ, ইদের কিনাকাডা করুন লাগবো ভাই।	ʃɐmne ɪd̪, ɪd̪er kɪnɐkɐdɐ korun lɐgbo bʱɐ͡i̯!

5.1.6 Voice or speech Diversity

During the development of this corpus, we aimed to ensure maximum diversity and representation in the speech data. We attempted to standardize the length of each sample to approximately 10 minutes. If a sample exceeded 10 minutes, the data collector was instructed to balance this by gathering additional, shorter samples from other individuals. Conversely, if a sample was shorter than 10 minutes, the collector was directed to obtain longer samples from other participants. Recognizing that vocal characteristics vary with age, we also endeavored to collect speech data from individuals of all age groups within each region through our assigned data collectors.

5.1.7 Gender Diversity

This speech corpus comprises data from a minimum of 183 speakers, including approximately 50% male and 42.7% female participants. Additionally, 7.303% of the samples feature multiple speakers from both genders, resulting in 89 male speakers, 76 female speakers, and 13 clips with multiple speakers representing both genders.

5.1.8 Geographical Diversity

Although a single region may encompass multiple dialects [62], data collection was conducted across various subregions within each region to capture the diversity of dialects and ensure geographical representation. In total, 31 subregions were covered across

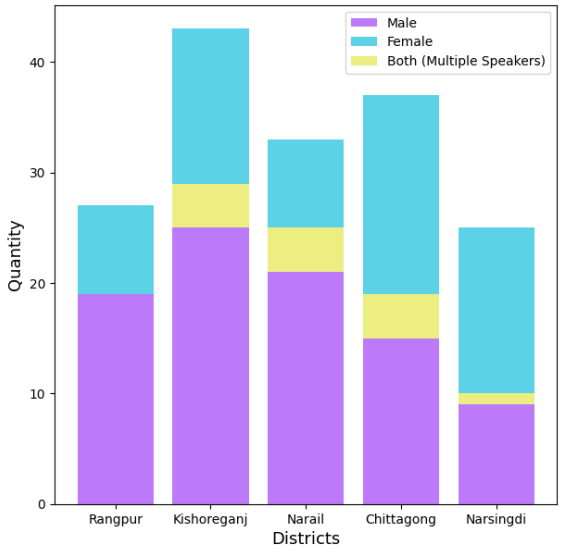

Figure 5.3: Gender quantity in the Regional Speech Corpus

the 5 districts from which speech was collected, namely Rangpur, Kishoreganj, Narail, Chittagong, and Narsingdi. These districts correspond to 1, 3, 17, 7, and 2 subregions, respectively, as detailed in Table 5.5.

Table 5.5: Subregions covered from each district

Districts	Subregions
Rangpur	Kamarpara
Kishoreganj	Pakundia, Bhairab,Bogadiya
Narail	Barakalia, Chotta Kalia, Kartikpur, Mirzapur, Ramnagar, Joka, Baka, Patna, Chandpur, Uthali, Jogania, Kalabaria, Pahardanga, Bil Bauchh, Joypur, Baidhyarbati, Gachhbaria
Chittagong	Nalanda, Satkania, Patharghata, Rangunia, Potiya, Hathazari, Boalkhali
Narsingdi	Polashtoli, Madhabdi
Total	30

5.1.9 Topic diversification

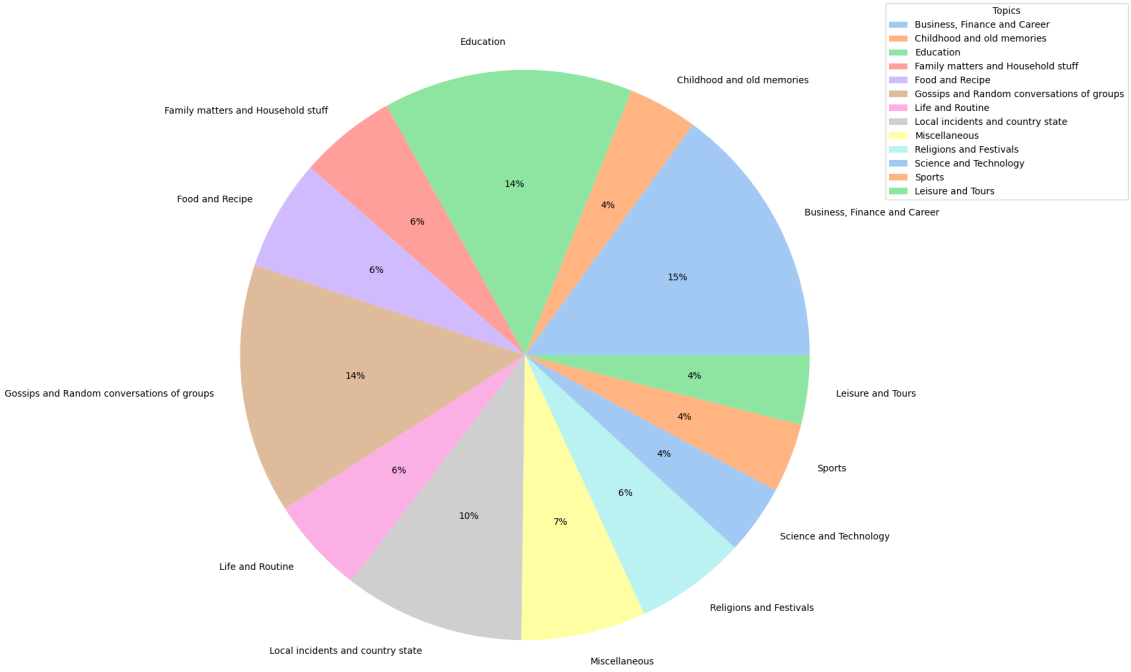

Figure 5.4: All subcategories of topics in the dataset

The dataset comprises speeches on 64 unique topics, categorized into 13 distinct categories. These categories are further organized into various subgroups of clusters, as illustrated in Figure 5.4.

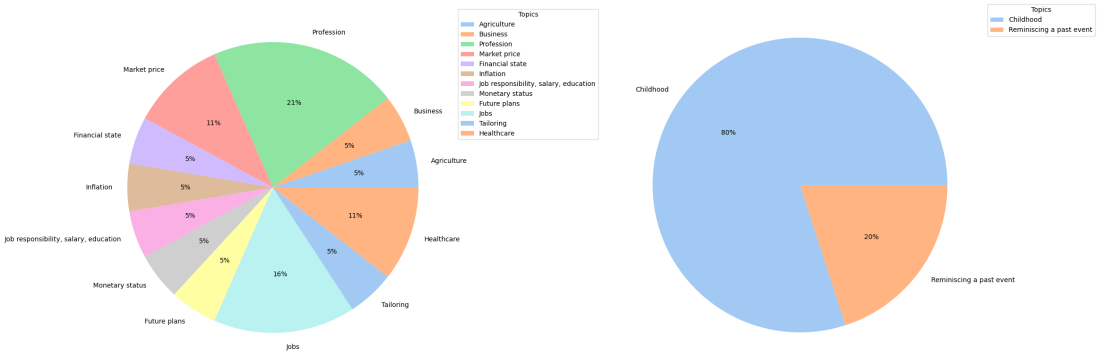

Figure 5.5: (Left) Topics in the ”Business, Finance and Career” subcategory and (Right) Topics in the ”Childhood and old memories” subcategory

5.1.9.1 Business, Finance and Career

This subcategory encompasses topics related to Business, Finance, and Career, as detailed on the left of Figure 5.5. The majority of the topics pertain to individuals’ professional

roles, job responsibilities, educational qualifications, and various professions through which they earn a livelihood. Given that a significant portion of the data was collected from rural areas, discussions frequently focused on agriculture, tailoring, and healthcare. Additionally, the subcategory includes topics such as market prices, inflation, financial conditions, and the monetary status of families or individuals.

5.1.9.2 Childhood and old memories

This subcategory encompasses topics related to nostalgia, as depicted on the right of Figure 5.5. It includes discussions about childhood experiences, past living conditions, and reflections on various events from the past. The majority of the conversations focus on childhood memories.

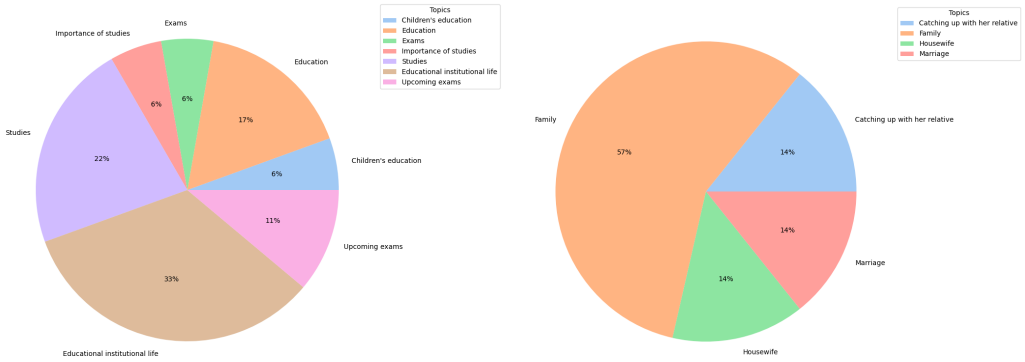

Figure 5.6: (Left) Topics in the ”Education” subcategory and (Right) Topics in the ”Family matters and Household stuff” subcategory

5.1.9.3 Education

This subcategory addresses topics related to education, as detailed on the left of Figure 5.6. It primarily features discussions by students about their examinations, academic institutions, and the subjects they are studying.

5.1.9.4 Family matters and Household stuff

This subcategory encompasses topics related to family and household affairs, as illustrated on the right of Figure 5.6. The discussions are predominantly conducted by older women and focus on familial matters and domestic issues within their households.

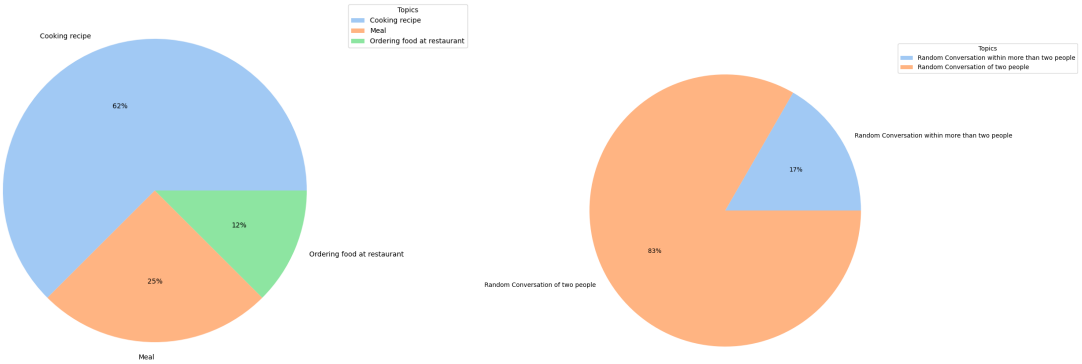

Figure 5.7: (Left) Topics in the ”Food and Recipe” subcategory and (Right) Topics in the ”Gossips and Random conversations of groups or individuals” subcategory

5.1.9.5 Food and Recipe

This subcategory addresses topics related to culinary matters, as depicted on the left of Figure 5.7. The discussions are primarily conducted by women and housewives, focusing on various recipes and food-related topics.

5.1.9.6 Gossips and Random Conversations of Groups or individuals

This subcategory primarily encompasses casual gossip or conversations among two or more individuals within a group, as illustrated on the right of Figure 5.7. The discussions often shift between topics, reflecting the spontaneous nature of the speech. The data collector was instructed not to interfere during these interactions. In some sample clips from this subcategory, there are more than two speakers, including the data collector.

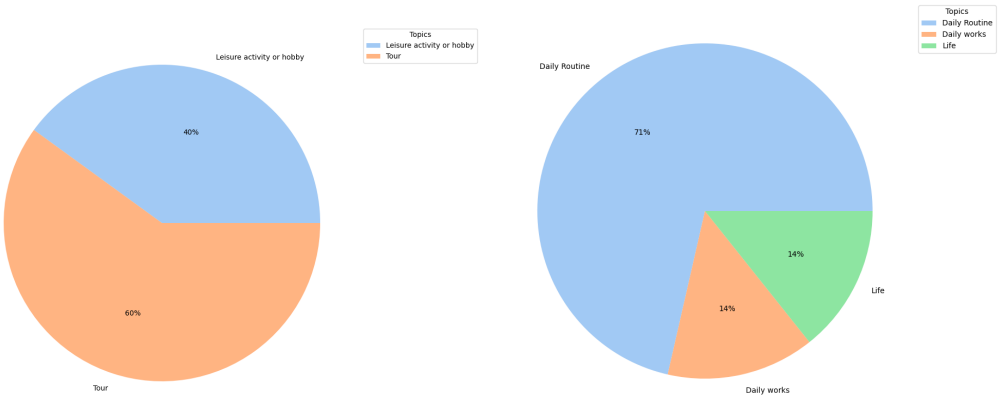

Figure 5.8: (Left) Topics in the ”Leisure and Tours” subcategory and (Right) Topics in the ”Life and Routine” subcategory

5.1.9.7 Leisure and Tours

This subcategory includes topics related to hobbies and travel, as shown on the left of Figure 5.8. The discussions cover various hobbies pursued by the speakers and the different trips they have planned or experienced with friends to various locations.

5.1.9.8 Life and Routine

This subcategory encompasses topics related to individuals' daily life and activities, as illustrated on the right of Figure 5.8. The majority of the speakers are students, and their discussions focus on the routine activities and daily experiences they encounter.

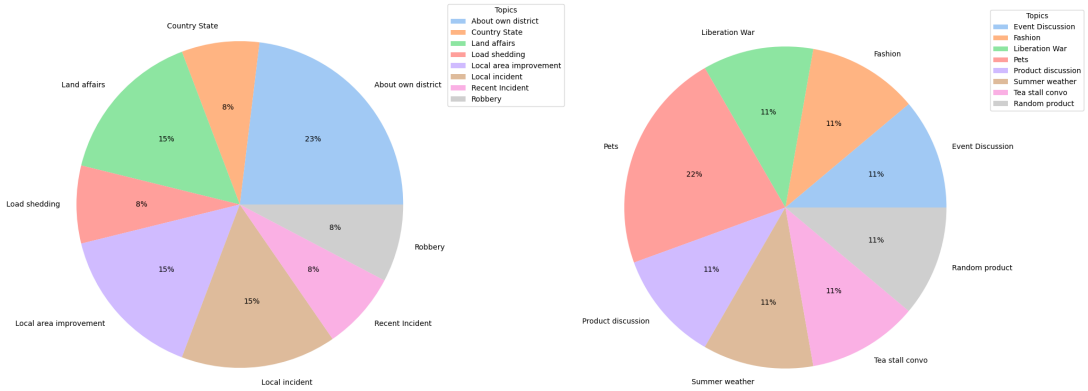

Figure 5.9: (Left) Topics in the "Local incidents and country state" subcategory and (Right) Topics in the "Miscellaneous" subcategory

5.1.9.9 Local incidents and country state

This subcategory addresses topics related to recent incidents occurring in the speakers' vicinity, as shown in Figure 5.9. The speakers, predominantly from rural areas, discuss various local issues, potential solutions, district-specific matters, local politics, and recent events within their communities.

5.1.9.10 Miscellaneous

This subcategory includes topics related to various miscellaneous subjects, as illustrated in Figure 5.9. The discussions encompass a range of topics, including different types of pets, fashion, and other diverse interests.

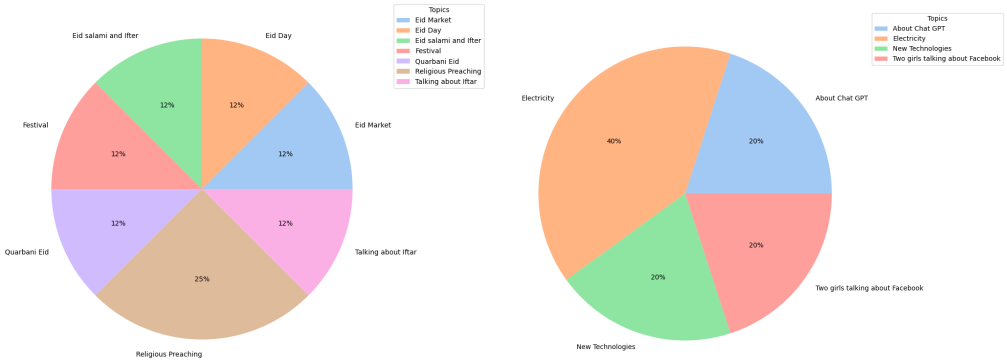

Figure 5.10: (Left) Topics in the ”Religions and Festivals” subcategory and (Right) Topics in the ”Science and Technology” subcategory

5.1.9.11 Religions and Festivals

This subcategory addresses topics related to religious activities. Given the predominantly Muslim population in the regions, the conversations primarily focus on Eid and associated practices. The topics and their distributions are presented on the left of Figure 5.10.

5.1.9.12 Science and Technology

This subcategory encompasses topics related to technology, as detailed on the right of Figure 5.10. Given that the majority of the data were collected from rural areas where stable electricity is a prevalent issue, discussions predominantly focus on this topic.

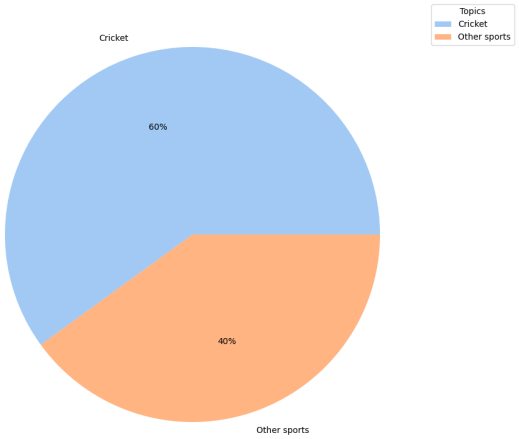

Figure 5.11: Topics in the ”Sports” subcategory

5.1.9.13 Sports

This subcategory includes topics related to sports, as shown in Figure 5.11. Given that cricket is the most popular sport in the subcontinent, the majority of discussions within this subcategory focus on cricket.

5.2 Exploratory Data Analysis and Feature Extraction

In this chapter, we delved deep into our developed speech corpus with regional Bengali dialects by doing some exploratory data analysis and relevant feature extraction and also illustrated how this regional speech corpus deviates from any standard Bengali speech corpus. We have used the OOD-Speech: A Large Bengali Speech Recognition Dataset for Out-of-Distribution Benchmarking [87] as the reference corpus for standard Bengali.

5.2.1 Exploratory Data Analysis

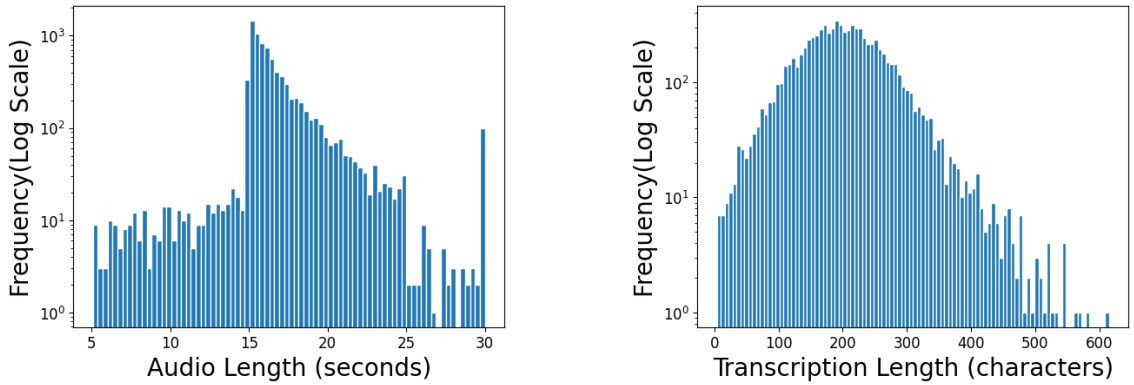

Figure 5.12: (Left) Audio length distribution of the regional Bengali dialect corpus (Right) Transcription length distribution of the regional Bengali dialect corpus

On the left of Figure 5.12 illustrates that the majority of the recordings have durations between 15 and 25 seconds, with a maximum length of 30 seconds. Additionally, on the right of Figure 5.12, it is demonstrated that there is no significant correlation between transcript character count and audio length.

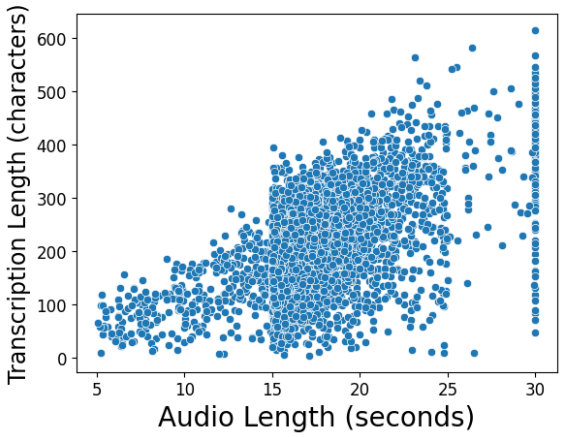

Figure 5.13: Transcription length vs audio length distribution of the regional Bengali dialect corpus

Although we do not observe any peculiar long transcripts for shorter audio recordings, the opposite is observed in several cases as we can see in Figure 5.13. Upon further investigation, these samples were found to be characterized by high levels of background noise, incomprehensible speech, or transcription errors.

5.2.2 Feature Extraction

5.2.2.1 Comparison with Standard Bengali

To evaluate the feature diversity within the Promito Bengali or Standard Bengali language dataset, we extracted Geneva speech features from $10,000$ samples from the referred standard Bengali speech corpus [87] and compared them with our regional Bengali dialect speech corpus.

Figure 5.14 depicts the long-term spectral average (LTSA) across different dialects. The analysis reveals that dialects not only diverge semantically but also spectrally. This observation indicates a significant distribution shift between some dialectal data and others. Given that the same recording protocol was adhered to, this shift is likely attributed to prosodic variations inherent in the different dialects.

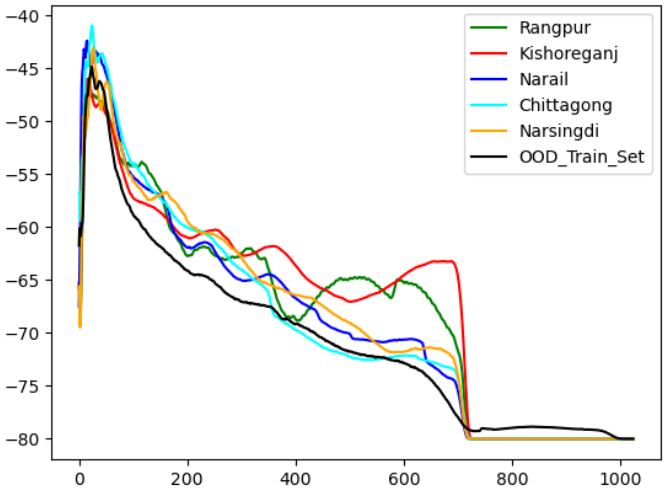

Figure 5.14: Long-Term Spectral Average plot of the regional Bengali dialect corpus

Additionally, we visualized the differences by plotting log-histograms of Geneva features for two sections: one representing regional district dialect speech data and the other representing Standard Bengali speech data. These comparisons encompassed various parameters, including pitch, loudness, and spectral attributes. The analysis revealed distinct variations in these parameters, as depicted on the left of Figures 5.15 for Bengali speech with regional dialects and on the right of 5.15 for standard Bengali speech. Each audio sample consisted of 88 features.

We further examined the distribution shift using Figure 5.16, which focuses on a specific feature, with 'ancholic' encompassing all the regional Bengali dialect audios and 'porimito' representing $10,000$ samples from Standard Bengali speech data. Figure 5.17 illustrates the differentiation between dialects within the regional Bengali languages.

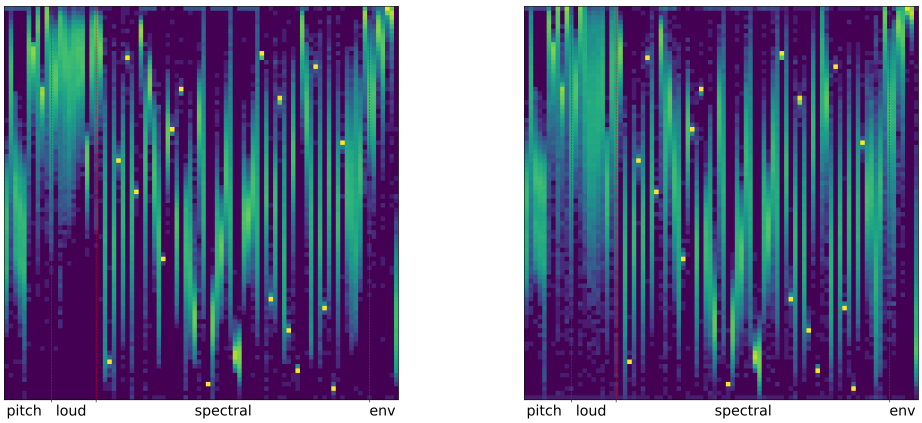

Figure 5.15: (Left) Stacked log-histograms of Geneva features for regional Bengali dialect (Right) Stacked log-histograms of Geneva features for Standard Bengali Dialect

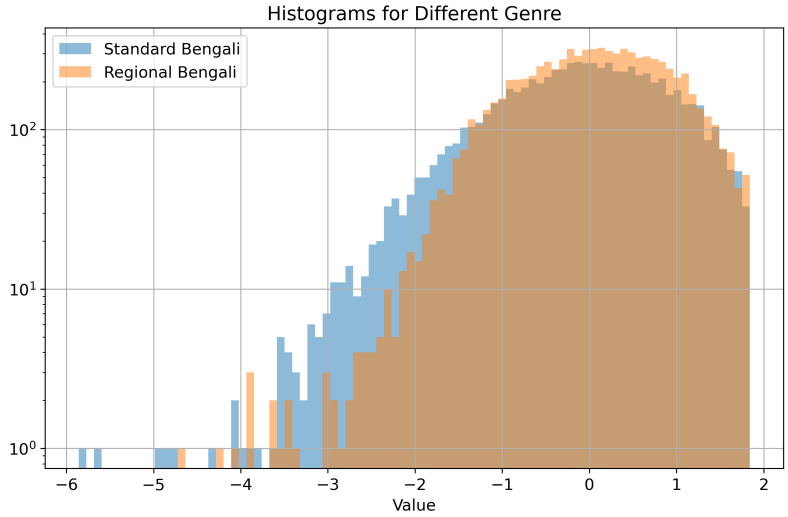

Figure 5.16: Histogram comparison between Geneva features of samples for Standard Bengali and Regional Bengali

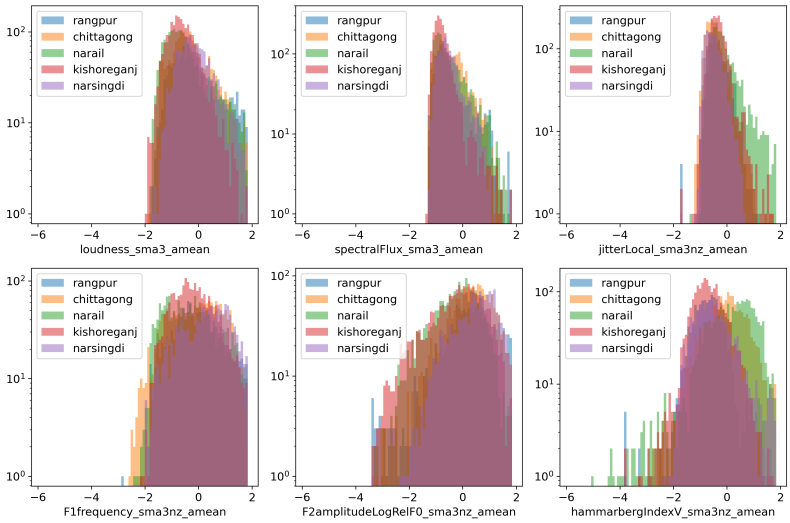

Figure 5.17: Histogram comparison between Geneva features of samples of different districts

To investigate the linguistic diversity of various regional dialects, we employed t-distributed Stochastic Neighbor Embedding (t-SNE) plots to visually represent the linguistic variation and distinct characteristics across different districts. By applying t-SNE to reduce the dimensionality of high-dimensional Geneva features and Wav2vec2 embeddings to two dimensions, we gained insights into the underlying structure of these dialect samples. The t-SNE plots based on Geneva features highlighted the distinctive linguistic characteristics within the Geneva dialect, allowing us to understand how these features cluster and relate to each other. In contrast, t-SNE plots using Wav2vec2 embeddings from various dialects provided a comparative perspective, enabling us to assess the similarities and differences between Geneva features and those from other dialects. The t-SNE plots of Geneva features extracted from five different districts in our dataset revealed clear separation between clusters, indicating significant differences in the Geneva feature space. A similar pattern was observed in the Wav2vec2 embeddings, with distinct clusters forming for different domains, as illustrated in Figures 5.18 and 5.19. These experiments were conducted using a perplexity of 40 and a 12-metric for clustering.

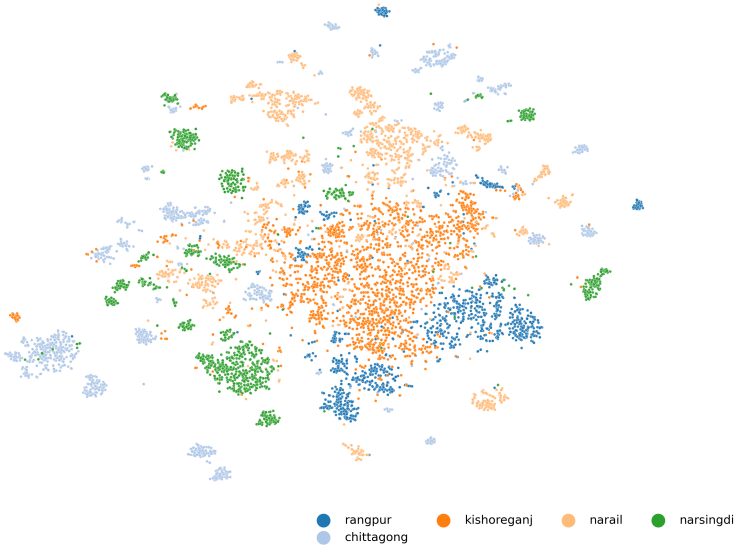

Figure 5.18: t-Schochastic Neighbor Embeddings of Geneva features of samples extracted from different dialect samples of the dataset.

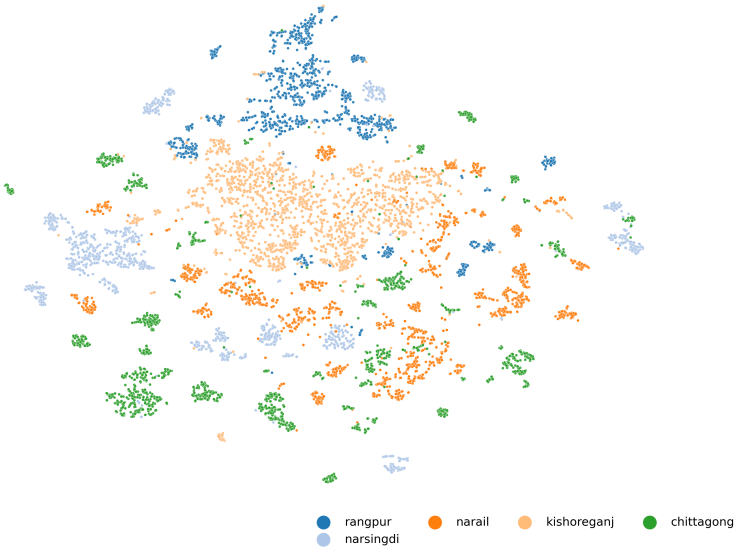

Figure 5.19: t-Schochastic Neighbor Embeddings of Wav2vec2 embeddings of samples extracted from different dialect samples of the dataset.

5.2.3 Spectograms

A spectrogram visually represents how different frequencies in a signal vary over time. It comprises three dimensions: time, frequency, and the amplitude of each frequency, with colors or brightness indicating the strength of each frequency. Spectrograms are

commonly used to analyze audio signals, identify sound patterns, and study signal characteristics. Spectrograms of random samples from the dataset were plotted to assess the quality and diversity of the data, as well as to visualize the differences among audio clips from various regions.

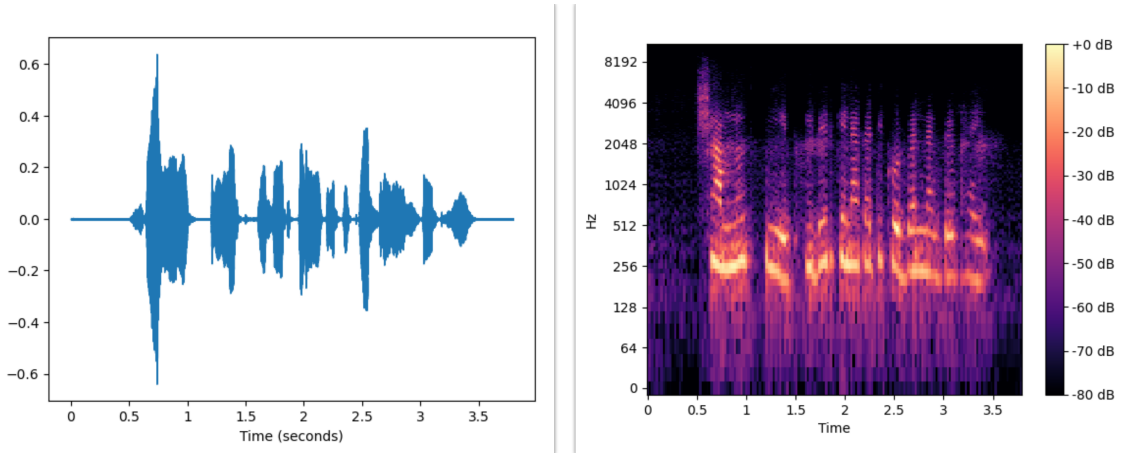

Figure 5.20: Waveform and Spectrogram of a sample from Rangpur

In Figure 5.20, a representative spectrogram for a data sample from the Rangpur fold is displayed. The spectrogram reveals that most temporal segments of this fold exhibit some level of energy, indicating fewer quiet periods. The high energy across a broad frequency range is evident from the bright regions, reflecting both the noisy nature of the recordings and substantial energy levels.

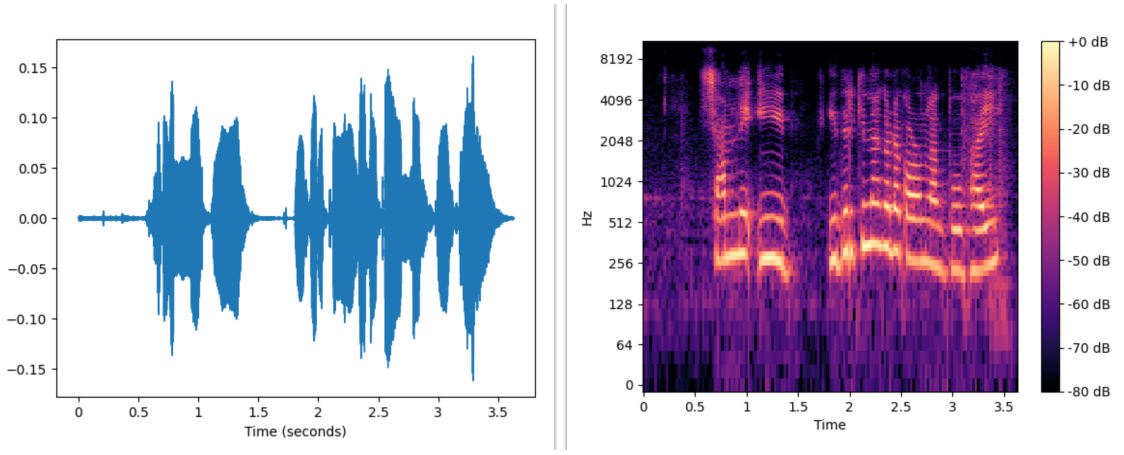

Figure 5.21: Waveform and Spectrogram of a sample from Kishoreganj

Figure 5.21 illustrates a spectrogram from the Kishoreganj fold, where the energy is clustered due to gaps between speech segments.

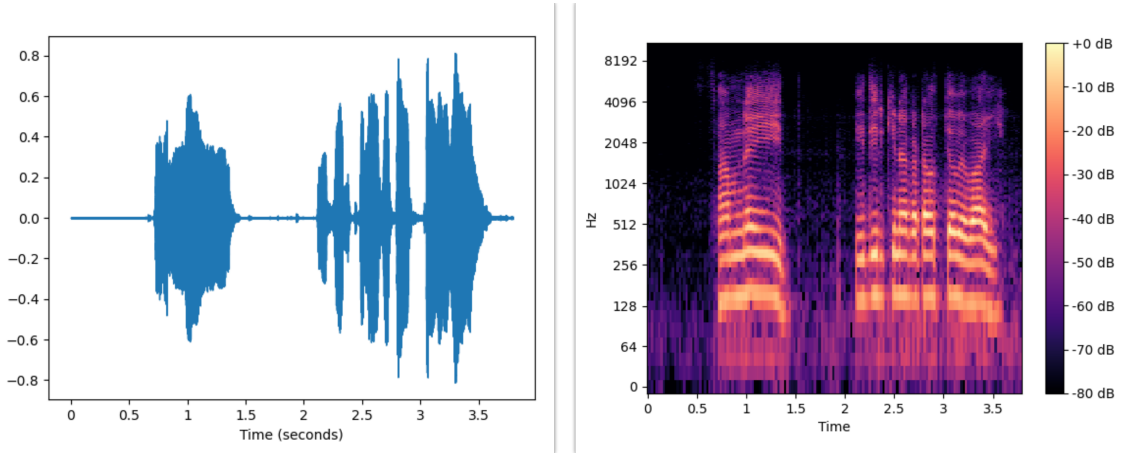

Figure 5.22: Waveform and Spectrogram of a sample from Narail

A similar observation is noted in the spectrogram from the Narail fold, shown in Figure 5.22, where the clustering of energy is more pronounced.

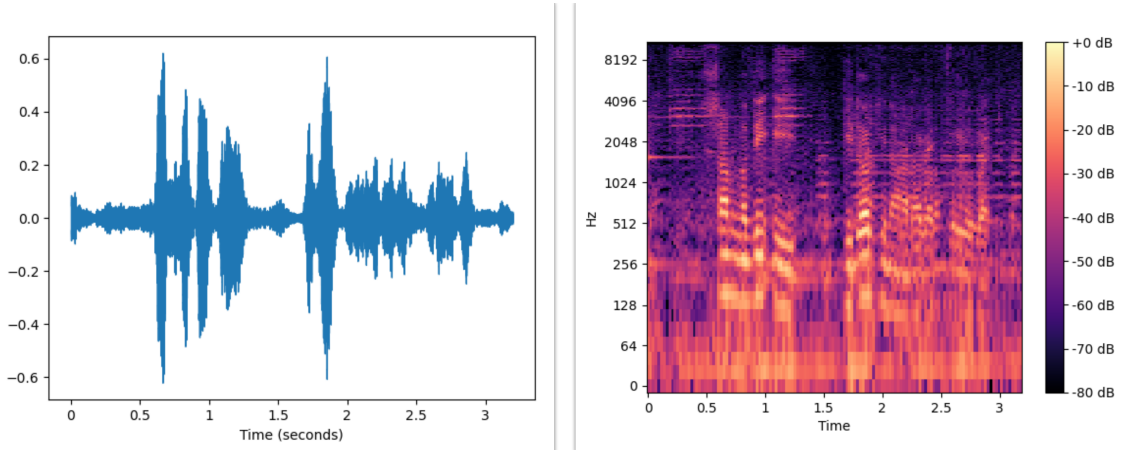

Figure 5.23: Waveform and Spectrogram of a sample from Chittagong

The spectrogram for the Chittagong region, depicted in Figure 5.23, shows significant energy, attributed to the emphasis or stress placed on particular syllables or segments by speakers from this region.

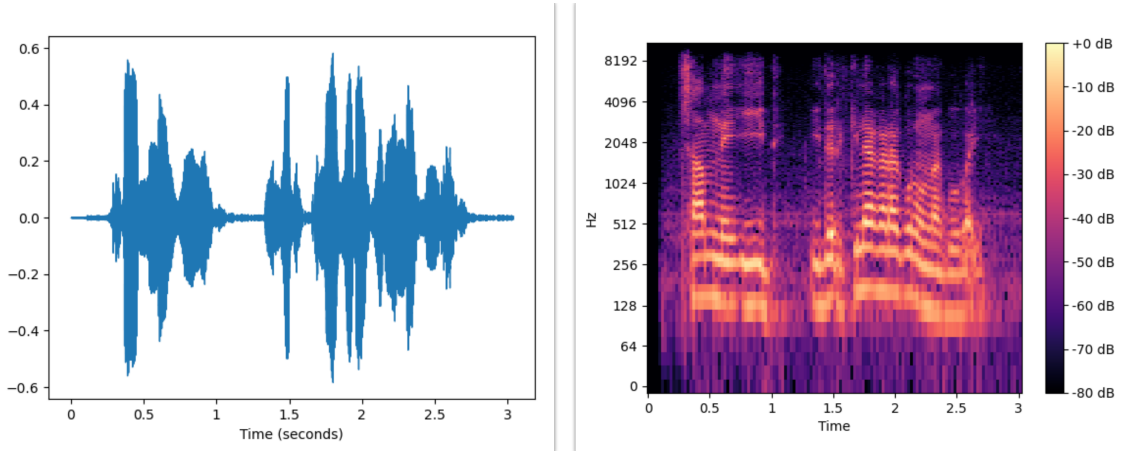

Figure 5.24: Waveform and Spectrogram of a sample from Narsingdi

Finally, Figure 5.24 presents the spectrogram for a data sample from the Narsingdi fold. The spectrogram indicates a consistent energy level across most temporal segments, with high energy over a broad frequency range, suggesting a noisy recording environment with substantial energy, as shown by the bright regions.

Chapter 6

Result Analysis

6.1 Evaluation Criterias

While fine-tuning the model, We evaluated model based on two evaluation criteria, Word Error Rate (WER) and Character Error Rate (CER).
The final WER and CER of this fine-tuned model are respectively 0.855 and 0.451. The table 6.1 shows District-wise WER and CER.

Table 6.1: Region-wise word error rate and character error rate

District	Word Error Rate (WER)	Character Error Rate (CER)
Rangpur	0.866	0.466
Kishoreganj	0.970	0.557
Narail	0.757	0.364
Chittagong	0.886	0.471
Narsingdi	0.798	0.399
Final	0.855	0.451

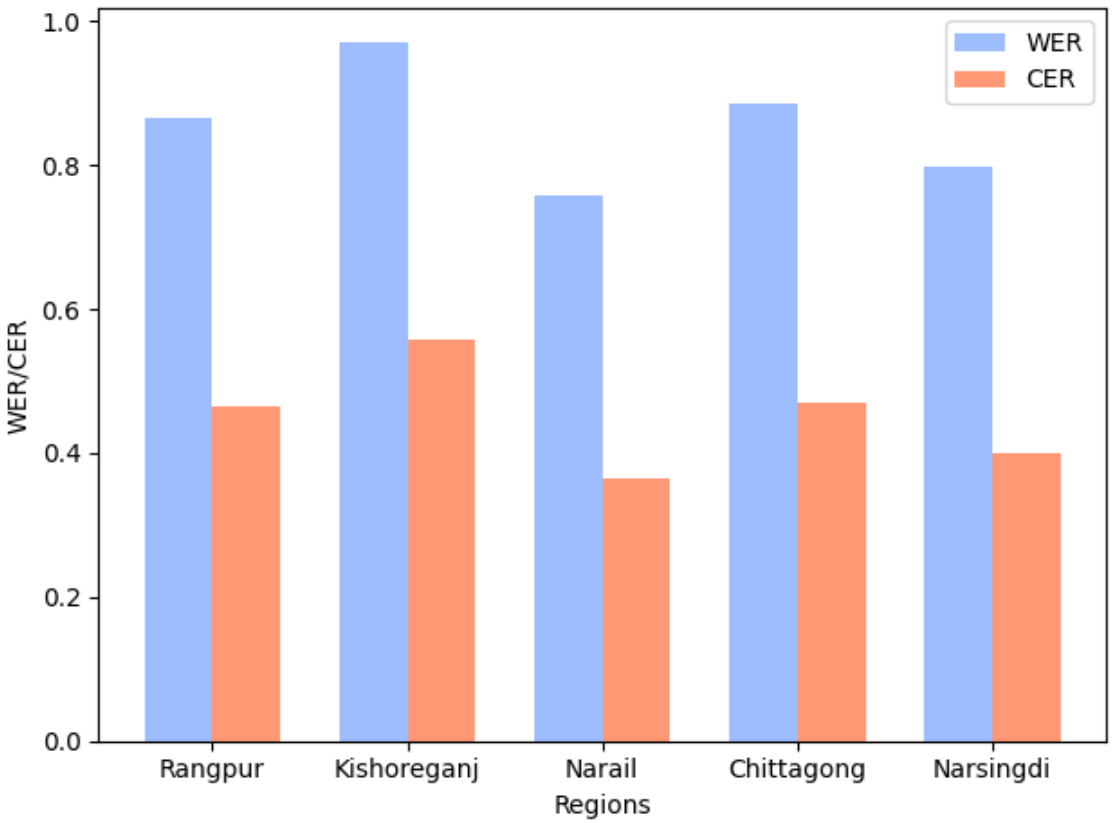

Figure 6.1: PX5's performance for each region

6.2 Model Inferences

Some region-wise inferences of the samples from our fine-tuned model are shown below in the figures 6.2, 6.3, 6.4, 6.5, 6.6, for the respectively Rangpur, Kishoreganj, Narail, Chittagong, and Narsingdi test data from the corpus.

A	B	C	D	E	F	G	H
file_name	External_ID	Region	Transcriptions	Predictions	Split	WER	CER
test_rangpur (1).wav	rec_22_big_audio_0_chunk_0.wav	Rangpur	লেখিয়্যা, কালি দিয়্যা, সেই সুন্দর করিয়া, এখন সেই সময়ে তো মনে কর, সিরাজেরা আছিল ব্রাজিল, ওয় ব্রাজিল সাপোর্ট করছিল। হ, একটা ফিল্ডিং গ্রুপ ছিলো। হ, ওরে কতা <> বুজচি বিষয়টা, ওয় তো এখনও বেরাজিল, বেরাজিলের সাপোর্ট করে। সিডায়, ওই সময় আমরা ওই কানির রুমটাত যকন আছনো একসাতে? যকন কী ইয়্যা হইছিল, ওই যে সুমন আর হইলো আলম <> জামাই, আর আমরা তো মনে কর তিনজন ছিলাম একসাতে যেমন, সিরাজ, আমি, তুই তিনজন ছিলাম, পাশাপাশি আমার সাতে	পলেহিয়া, কালি বিয়ে, দেখি সুন্দর করি।? এখন সেই সময় তো মনে কর সিহাজেরা আচিল ব্রাজিল, পর এটা কিনিনম করচিল, ঢেখা? পনে কত আটি ো এখ ্ো, ন না সাংজ কবাবা সেইটায়, ওই খা ওইযে, আমরা ওই খালি রুমটের যোকোনো আচিলো এক সাতেুম, যকন কী ইয়া হচছিল। এবাগে সুমন আর হইলো আলম খালি র ঘ্যাঞ্র জাময়। আর আমজা তো মনে কর কীন জনছিলাম একসাতে। যম শ্িব, আমি তুই নভা ডলা, পাশা-পাশি আমার সাতে	Test	0.75	0.452322
test_rangpur (10).wav	rec_23_audio_5.wav	Rangpur	হারাগাছে? হুম! <> তাইলে ইয়া, জমি জমাত তো আবাদ করেন নাই, না আপা? <> না মোর জমিন নাই, ও গরু কয়টা আছে? গরু আছে একটা। হাঁস-মুরগিও আছে না? হাঁস-মুরগিও আছে।	খারা কছেন? পরেখ না মুহি। তাইলে ইয়া জমি জমার তো আবাদ করেন নাই। না আফা? না মোর জমির নাই। গরু কয়টা আছে? গরু আছে একটা। হাঁস-মুরগি আছে না? খাস-মুরগিও আছে । ও	Test	0.633333	0.24
test_rangpur (100).wav	rec_24_audio_39.wav	Rangpur	মুই কেনা কাজলি মাছ খাবার পাও না। ক্যা? কাজলি মাছ তো ভালো। মোক অত ভালো নাগে না। তোর খামা খামা কত সুন্দর। খাইতে মজা। মোর কাছে কলম্বো গুলেয় ভালো নাগে। হামার আনলে তো কাজলি আনে। মোর কাছে কলম্বো গুলেয় বেশি ভালো নাগে। <>	মুই কেনা কাজুলে মা ক কাজনী মাছ তো ভালো? খামা কামামাটতো সুন্দর। খাইজে মাজার। ান কমদাট টু ভালো নাগে। মে ত কাজ আনে। ু ক কা তে নি তা ভালো নে	Test	0.790697	0.502347
test_rangpur (101).wav	rec_24_audio_40.wav	Rangpur	তোমার উকুনে কামড়ে কামড়ে ঘাও পচড়া বান্ধায়।এবারে <> মাছগুলে পাওয়া যায় না তোমার এড়ে? এখন পাওয়া যাইবে? পানি যখন হইবে তখন পাওয়া যাইবে। দুই একটা বৃষ্টি হইলে। তখন পাওয়া যাইবে। না? হুম।অনেকের ডিগির মাছ বারে আইসে তখন সেই সময়।	তমারে ে হ রে ঘাও পচ্চে নে। এলামাছগুলা পাওয়া যায় না তোমার এড়ে? এখন পাওয়াবিলে, ঝকন হইবে তখনএ দুই একটার ডিষ্ট তখন পাওয়া যাবে না? অনেকের ডিগির মাছ বাড়ে আইসে তকন সেই সময়।	Test	0.666666	0.398148
test_rangpur (102).wav	rec_24_audio_41.wav	Rangpur	দোলাত পানি আটকি থাকে না বিল গুল্যা? হুমনিচের বিল গুল্যা? ওই মাছ গুল্যা নামি আইসে। অড়ে থাকি বালিয়া পাওয়া যায়। নোয়ায়? মেলা করি ইয়ের আগে যে বালিয়া খায়া আনু একদিন। ট্যাংনা, বালিয়াএ বৃষ্টি হইলে দারকি বসাইলে হামার খাওয়ার মাছ হয়।	দুলাফ পানি এটকি থেকে না বিলগুলের মানে। বিলগুলে? হুম।িচে পেরে ওই মাছল মানশি। ই যে বাড়িয়াও পাওয়া যায়, নয়া, মেলা পড়ি ইয়ের আগে যে বালিয়ে খায়া নয়দিন ট্াং না বালিয়া? এ বৃষ্টি হইলো আর কি বসালে হামার খাওয়ার মাজ হয়।	Test	0.85	0.427927
test_rangpur (103).wav	rec_24_audio_42.wav	Rangpur	মাঝে মাঝে খায়েন <> । ভম ভম ভম বোও। হায়রে মেলা রে মেলা। মেলা তাউক যেতে আশুপতন গেইসে মেলা। ওই মেলাত গেইলি ওয় খালার বাড়িতে।	মাজে মাযে খাই আমরা চর ছ না খা। ধম দাম্বামদা, মলাবে মেলা। মেলা তাগ হাসু কত গেইসে মেলা? ত ওই মেলাত গেইলে ওইরার খালার বাড়ত যেয়।	Test	0.84	0.475
			আইজকা বলে মেলা? হামরাও তো শুনবার লাইগসে মুই রোজা রাখসে, রজার মাসে, চল যাই ঘরে । মেলা হইলে ক্যামনে কী হয়? কী করবে মেলা যাইয়ে? আইজকা যাইয়া খাওয়াও যাবান না, ঘুরাও যাবা<> ঘুরবার গেলেও ওটে কইলজ্যা শুকাই যাইবে। আমরা তো ওইযে গেলে ওটে গেসিলাম। <> না হইলায় যখন	আইজপা বে হলমলে তো শুনবার লাগরজা হটে টনওে। মেলা হইলে কোমনে কী হয়। কী কইরবের মেলার যেবে আইজকে যে খায়া যটাটা গুরা যিপে া দ্যা সুচে বিচে দুরবার গেইলে ওতো কইলচে সুবি যাইবে। হমরা তো ওই যে গেইলে ওড়ে গেসনো।			

Figure 6.2: Model inferences samples on Rangpur data

A	B	C	D	E	F	G	H
file_name	External_ID	Region	Transcriptions	Predictions	Split	WER	CER
test_kishoreganj (1).wav	rec_41_audio_20.wav	Kishoreganj	<> এইযে ওয়া। <> ছবি তুলছস? তে আস না ক্যা? দে	া হুম, তুতে তর ফারে।ই? ত ত নাইগা?রে	Test	1	0.659090
test_kishoreganj (10).wav	rec_41_audio_29.wav	Kishoreganj	মানুষ অহন ঘুরাঘারির মইদ্দে আশুত্তিয়া বেশি, নাইম? কই, তুই <> মালারের ছবি ইয়া। <> দিয়ালতাম?	মানুষন ঘুরা গারির মইদ্দে আশযক বেশি নো?া কই? হেযে মাজার ছবি, ইয়াই মব্যা টেহালা	Test	0.933333	0.449438
test_kishoreganj (101).wav	rec_42_small_audio_1.wav	Kishoreganj	<> গাছ কি এমুন? না অইন্য কালার? না, এমনই।	আইয়া দি গরমাঝে। এ হেই দেশে খেজু গছ কি মু না অইন্য কেলা? না মুনই।	Test	1.333333	0.975609
test_kishoreganj (102).wav	rec_42_small_audio_2.wav	Kishoreganj	মুস্ক কি? কয় কিয়া বাত। কিয়া বাত। কি?	মুস কু কি? ক কি আবাদ? কি আবাদ। কি?	Test	0.875	0.333333
test_kishoreganj (103).wav	rec_43_audio_0.wav	Kishoreganj	ক্যারে ফরতাছো? এনতে এই জাগাত কইছে না?হুম।ফর,ফর, ফর। উদ্ভিত তার খাদ্য বৃদ্ধির পরাগায়নে ও জিবের বিস্তরের জন্য প্রানির উপর নির্ভরশিল।নির্ভরশিল টিকাছে। উদ্ভিদ কুনডা আর প্রানি কুনডা এইডা ক?	ত। হুম। এত এইএে কইছই না? এফরফ ফ উদবিক থার, খাইদ্দারতে এইডে বিদ্ডি পড়ানাড় ন, অদদিবের বিছতারর জন্য? পানির উপর? নিরবসী। নিরবশীল, ঠিকাছে। এ উদ্দিশকুন্যা আর ফানি কননে ইযে ক।	Test	1.035714	0.538043
test_kishoreganj (104).wav	rec_43_audio_1.wav	Kishoreganj	উদ্ভিদ অইছে গাছ-পালা। প্রানি- গরু, ছাগল, মানুষ।উদ্ভিদ অইছে গাছ-পালা, প্রানি অইছে? গরু, ছাগল, হাস-মুরগি, মানুষ।গরু, ছাগল, হাস মুরগি টিকাছে? হ্যাঁ। মাত্ উদ্ভিদ তেকে বিভিন্ন স্থানে ছড়িয়ে এই যে, এই যে জিবে ছড়িয়ে পরাই হলো বিজে বিস্তরন।বিজে? বিস্তরণ।বিস্তরণ।	উদবিদি। গাছপালা প্ানীব। গরু সাগরে। ভর রে ইসে গাছ পালা প্যানি অইছে, গরু সাউগলবা। পন মাশমুষী, ঠিকাছে? হ্যাঁ। এইযে মাতরী দ বিত্তেকে বিভিন্নসানে ছরিয়ে হৈ জিবে ছরিয়ে পড়ার হলোব বিজে? বিস্তরন বিজেশর? বিস্তরন, বিজে?	Test	0.972972	0.484126
test_kishoreganj (105).wav	rec_43_audio_2.wav	Kishoreganj	বিস্তরন ঠিকাছে?আই ফানি খাইয়া ল। ইয়া বই ল,সাধারন গ্যায়ান বই ল। এইডা আর কয়ডা। এইডা তো কয়ডা তো ফরছছোই। ফাচ আর তিন কতো? আট। চাইর আর এক? তিন আর এক চাইর। ঠিকাছে। শুন্য শুন্য। হ্যাঁ অইছে যাও সাধারন বই লও।	বিস্ত টিক ছে? া সাটি ইয়ও অইল, এা বইল, সাজাহঙ্গেন বইল। কয়দে এইরো তো কয়ডা তো পাছুছুইজ? পাচ চার্নিন কত আট। চাইর আইলে টিনারে? চাই। টিক আছে। সুন্নশনে, হ্যাঁ, ইসলা ইয়া সাজাউন বইল	Test	0.9	0.477157
test_kishoreganj (106).wav	rec_43_audio_3.wav	Kishoreganj	আগে আমারে অংক দেও। পরে সাধারন গ্যায়ান বই। এইযে, জিবে, জিবে বিস্তর নতুন উদ্ভিদে, নতুন উদ্ভিদকে আবার গরে তুলতে সাহায্য করে। এভাবেই পরিবেশের উদ্ভিদ ও প্রানি একে অপরের উপর নির্ভরশিল। এভাবেই পরিবেশের উদ্ভিদ ও প্রানি একে অপরের উপর নির্ভরশিল।	আহযা জিবে, জিবে বীস্তার নতুন উদ্বিরে,নতুন উদ্বিরে আবাশ গরি তুলতে সাহায্য করে। এবাবে পরিবেশে উদবিতর পানি? একে প রেন হেখে। এগরদ্দে আদবি পনি একেপর	Test	0.820512	0.506382
test_kishoreganj (107).wav	rec_43_audio_4.wav	Kishoreganj	হাত দো গা বাইরে যাইয়া।কতো জানি ক? তিন কদম।হেইদে, পুকা, মাকর, উদ্ভিদ কেয়ে বেচে তাকে, পুকা, মকোড় উদ্ভিদ কেয়ে বেচে তাকে, পুকা, মকোড় উদ্ভিদ কেয়ে বেচে থাকে। এইযে আমারও আবার, পুকা, মকোড় উদ্ভিদ কেয়ে বেচে তাকে। আমরা এইযে আবার।	ছান্দকারইর দিলে। তিন েইমান। হেযে, পুকামা কর উদবিথকে বেইচা থাকে পুকামা কর উদবিথকে? বেচি তাকে। পুকামা কর উদবিথকে বেচি তাকে এই যে আমাযে আবার।	Test	0.948717	0.568807
test_kishoreganj (108).wav	rec_43_audio_5.wav	Kishoreganj	এই এইনি একটু ফরোছেন। শুইনা শুইনা একটু ফরোছেন। কুনে? এই একটু ফরোছেন। সকল প্রানীর শক্তি প্রত্যক্ষ বা পরোক্ষ ভাবে উদ্ভিদের উপর নির্ভরশীল। হুম	এহনো কটু পজান? টু পজান। কুনে? এই ইটু ড়ছেন। সকল প্রণের স্তি, প্রত্তকক্ষবাপ,্ পড়ক্ষবাবে, চুদ বিদের উপনির বচনসম	Test	0.869565	0.492753
			উদ্ভিদ সুর্যের আলো ব্যবহার করে <> নিজের খাদ্য নিজেই তৈরি করে। পোকামাকড় উদ্ভিদের, উদ্ভিদ যে খেয়ে বেচে থাকে। আবার	দবিচুচুরযের অলবে গর গে নি টু খাদ্যে নি যেই ত বোকামাকর,			

Figure 6.3: Model inferences samples on Kishoreganj data

A	B	C	D	E	F	G	H	I
file_name	External_ID	Region	Transcriptions	Predictions	Split	WER	CER	annotator
test_narail (1).wav	rec_32_audio_33.wav	Narail	বাগাইনে বাছাড়, ফুইলেসশর, নৃপেন বাছাড় আর ওই যে এট্রা কি নাম জ্যানু। ফুইলেসশরের ভাইর বিটা। বুইছো? কি ভালো নাম। ফুইলেসশরের জমি তাও তো তোর বাবা এই টাহা-ফয়সা দিয়ে।	পাগ্য-নিয়ে বাচার, ফুইলেসশর নিরিপেন বাচার আর ওই যেট্টা কি নাম জানু ফুইলেসশরের বাইর বিইটা বইছি? কি ভানো না। দিশশরে জমে তাতে তোর বাবা টা ফয়সা দিয়ে	Test	0.75	0.303797	aponsharif111@gmail.com
test_narail (10).wav	rec_32_audio_42.wav	Narail	আমি বাবুরে চিগিসসা তা ও কইছেলো কি যে বাবুরে চিগিসসা অন্তি টাহা দিছিস। তা বৌদি তো দশ হাজার টাকা পাইছে। আমরাও তো জানি দশ হাজার টাকা বৌদি পাইছে। দশ হাজার টাকা। তা কচ্ছে কি যে না আমার মাইসোতো ভাইর কাছে আমি অনেক টাকা দিছি। ও নিজি কইছে। এওন যদি কয়ে যদি কয় যে আমি কই নেই।	আমি বাবুরে চিগির সাতা ও কইছেলো কি যে বাবুরে চিগির সাতা টা দিছে ত বউ দি তো, দশ হাতা টাকা পাইছে আমরাতো জানি দশ হাজারটাকা বউ দি পাইছে। দশ হাজার টাকা। তা কচ্ছে কি যে নায় আমার মাইনষোত বাইর খাছে। আমিও অনেক টাকা দিছি। ও নিজি কইছে ও যদি কইয়া যদি কয় যে আমি কই না।	Test	0.482142	0.178030	aponsharif111@gmail.com
test_narail (100).wav	rec_34_audio_22.wav	Narail	ওই, বিপুল বাবু বিদেশে থাকতি তুই আর আমি খুলনাতে ডাক্তার দেখাইছিলাম মনে আছে? আড়াইশো বেডের তে? আমার থাইরয়েড ধরা পড়িছেলো। ইন্ডিয়ায় যাইয়ে আমি। আমার বুক কাইপে উঠিছে জানো? ধরলাম না ছুন্নাম না আবার ডাক্তারের কাছে। ইন্ডিয়ায় যাইয়ে ডাক্তার দেখাইছি আমার থাইরয়েড নাই।	ওই বিপুল ভাবো বিদেশে থাকতি। তুই এরে আম খুলনাতে ডাক্টরদেয খাইছিলাম মনে অইছে। আড়াইশবরের্তে। আমার ছাইরের ধরা পড়িছিলো। ইন্ডিয়া হির। আমার কগই। মই ব খালে ডিতে জানো। ইন্ডিয়াই বা ডাকতে। ডর-লাম না চুলাম না আবার ডাকতারে আছে। ইন্হিয়াই ে। ইন্ডিয়াই যে। ডাক্তার দেয খাইছি আমার ছাইরা উয়ের নাই।	Test	0.880952	0.429133	razoanislamsimanto78@gmail.com
test_narail (101).wav	rec_34_audio_23.wav	Narail	এখন। যাক আলহামদুলিল্লাহ। কিন্তু কঠিন রোগ বইয়ে নিয়ে বেড়াচ্ছি এ আর কউয়ার মতো না। যত দিন থাকতি হবে অষুধ খাইয়ে যাতি হবে। কী? কী ম্যাডাম?	জাগালহামদুলিল্লাহ। কিতু কঠিন রোগ কিন্তু কঠিন রোগ বইয়ে নিয় বেড়াচ্ছি। এ আর কয়ার মতোর না। কিহৈ যাইতোদেন থাকতে বসুদ খাইয়ে যাইতে হবে। কী অবে না?	Test	0.76	0.484848	razoanislamsimanto78@gmail.com
test_narail (102).wav	rec_34_audio_24.wav	Narail	বুকে টিউমার ধরা পড়িছে না! মানে বুকে বলতি? ভিতরে? হ্যাঁ । অপ্রেশন করিছো? অপারেশন করার কথা কই নেই? না, না। টিউমারটা এমন জায়গায় সেটা তোর	আম বুকেট উনার ধরে পইরে জানেই। মানে বুকে বলতি ভিতরে?য। চা অপারেশন করিছো? মানে অপারেশন করার কথা কই নি? না না টুইমারটায় আবন জায়গাই সেটা দর।	Test	0.76	0.338345	razoanislamsimanto78@gmail.com
test_narail (103).wav	rec_34_audio_25.wav	Narail	আমি যে ডাক্তার দেখাইছি ওষুধ খাওয়ার পরে অনেকটা <> ভেঙে ছোটো হইয়ে গেছে, বুঝছিস? অনেক সুময় থ্যালাসেমিয়া হয়, মানে মাংসপিণ্ড আরেট্টা হচ্ছে টিউমার। না মাংসপিণ্ড না, আমার একবারের জাগা ছয়বার	আব আমি যে ডাক্টার দেখাইছি। ওষুদ খওয়ার পুরো অনেকটা ই টিউমাট্টা বেহেগে ছোট হইয়ে গেছে। বুচছিস? না, অনেক সুমা যে ওই যে থেলাস মিয়া অয় মানে মাংস-বিন্ড। আরট্টা চে তী বাতি পানুন। আমার এক বেডা জাগায় আমার ছয় বাররে	Test	1.133333	0.532967	razoanislamsimanto78@gmail.com
test_narail (104).wav	rec_34_audio_26.wav	Narail	ই করা হইছে, ব্রেস্ট ইউএসজি হইছে ছয়বার। কয়দিন আগে একজনরে নিয়ে আসছিলাম, ওর ও টিউমার হইছেল ঘাড়ে। অপারেশন করিছে ঠিক হয়ে গেছে। তোমার ওইডে অপারেশন করার সিস্টেম নাই? অপারেশন করার সিস্টেম আছে কিন্তু	আকরা ওইছে। বরেস টিবাজজি করা অইছে ছয় বার। কয়দদিন আগে একজন একজনে নিয়াসছিলাম ওর টিউমারইছেলো ঘাড়ে। অপারেশন করিছি ঠিক ওইয়ে গেছে। যা তুমার এটা অপারেশন সিস্টাম নাই? অপারেশন করের সিস্টেম আছে। কিন্তু	Test	0.757575	0.310526	razoanislamsimanto78@gmail.com
test_narail (105).wav	rec_34_audio_27.wav	Narail	এখন তোর দাদা আগের দিন বলতেছিল যদি অপারেশন করতি হয় তাইলে ভেলোর চইলে যাবে অপারেশন করতি। কিন্তু, অপারেশন আপাতত করা লাগতিছে না। ওইডে ভেঙে ছোটো হইয়ে গেছে আর আগে যেমন জ্বালা যন্ত্রণা ই-টি করতো না? সেরকম করে না বুঝছিস?	এখন টো দাদা আগের দিন বলতেছিলো যে যেদি অপারেশন কইতি হয় তালি বেলেস চলে যাবা অপারেশন করতি। কিন্তু অপারেশনা পারেছত করা লাগতিছে না। ওত ভ্যেইঙে ছোট হইয়ে গেছে আর আগে য্মন জ্বালার জম্প্রেনা ইটি করতো না? সেরকম করে না বুঝছিস?	Test	0.538461	0.180094	razoanislamsimanto78@gmail.com
test_narail (106).wav	rec_34_audio_28.wav	Narail	যাক, যাক। ওইটা হচ্ছে গিয়ে যখন ডাক্তার দেখাইছিলাম ইনফেকশন হইয়ে গেছিল। আর এবার দেখালাম তা ইনফেকশনের ইটা কইমে গেছে আর ওটা ভেঙে গেছে, খণ্ড-খণ্ড হয়ে গেছে। যাক।	যাগোবাড়। এম ওইটা ইছে গে যখন ডাকটের দেখাইছিলোম। ইনফ্রেকশন হইয়ে গেছিলো। আর এবার দেখালাম তায ইনফ্রেকশনের ইটা কমে গেছে। আর ওইটা ভেইঙে ছোট হইয়ে গেছে। বনম খন্ড ভান্ন হইয়ে গেছে।	Test	0.851851	0.389610	razoanislamsimanto78@gmail.com
			<> সংবাদ। আর, আছে শুধ ফ্যাট টিস্যটা, ওটার এখনো কোন					

Figure 6.4: Model inferences samples on Narail data

A	B	C	D	E	F	G	H
file_name	External_ID	Region	Transcriptions	Predictions	Split	WER	CER
test_chittagong (1).wav	rec_29_audio_88.wav	Chittagong	গ্রামর মানুষ অল এইলে। মারেম্মা। ইন চায় দে ইতারা। হারণ ইন মাতি আইসসি দি এহন ন মাতিলেও নিজরতুঅ হারাফ লায়। <> ন মাতিলি। তারফরো আঁর উইবার লয়ও এনে এতিক্কিন হইজ্জা ঝগড়া <> আঁর লয় কি এঁডে? আঁর লাই তো বেককুনে ফরান দি ফেল্যায়।	গ্রামর মানুষ হল ইল্লে গের তেব মারেথা ভ লো ইতারে। অ্যারনইরে মাতি আইসসি দে এহন ন মাথিলে নিজ্ঞি খারাপ লাই বুঠিয়রে ন খরের উটে লয়নে তিককিন খতাইলে আআর লয় কি এডে? লা, আরলা আছে বেয়াকগুন ফারেন দি বেডে	Test	0.7906976744	0.472222
test_chittagong (10).wav	rec_29_audio_97.wav	Chittagong	<> ঢেইলি হওয়ার দরে বাইঝতো লাইগগো, হিসাব গরে, তুই হিসাব গরো না, এন্ডে হনে কি হর নে! ওরাইয়া যাই ফেলার নে হনো স, না না ওবা <> খালা আছেদেই, আই তো এনডিলালা নদি তদি ইদানিঙ।	খ বে না? হয় টেহ ঢেলে ফউয়ের ধরে বাইরযেলাই গসাব গরলো। ফুই সাব গরো না এনে কি, খনকের খন না? ঘুরা ইয়ে খাইবা লরনি খন স। না না বয়া দুরেইয়েরে খালাই আসে দে খালা, আঁই শইল্যে ন দি ট দিয়া দ	Test	1.085714286	0.662650
test_chittagong (100).wav	rec_31_audio_43.wav	Chittagong	আঁর ইংলিশত টার্গেট গজ্জিলাম এবারে 'এ-প্লাস' আইবু। 'এ-প্লাস' তো দূরত তাক, 'এ' ও ন আইয়ি, 'এ-মাইনাস' আইয়ারে বই তাইকি। আইণ্ড বাই ওইরইমমে ভাববিলম, এহন আঁর তে আইস্যে দে অইল দে 'সি'। তো আঁই মনে গইজ্জি অয়তো ইংলিশর মইদ্যে হন ভুল ইয়া টিয়া অইয়্যি।	ইর ংলেশি কাটিটজিলাম এগরে এ প্লাসসাইবু, এ প্লাস তুর দুরা থাক ও না আইয়ি, মাইন্রে সাইয়েরে বই থাইক্কি। তো আমি, আঁইণ্ড ভাইয়া এরিমশুইবিনাথারা আঁর তুই আইসসে দিসি।ই তো হআইমনে হজদ্যে ইগলিশ মলো খোনেইয, ভুলল্লি-টিয়া অইয়িন	Test	0.9069767442	0.527196
test_chittagong (101).wav	rec_31_audio_44.wav	Chittagong	আবার কেন গইজ্জেদে, <> দেইহিদে সেভেনত সেইম, নাম্বার নলরে চরে, ও আইচছা। তগগই, কি আর গরগুম, কুয়লাত যিয়ান আছে ইয়ান লইততাই ইইততাই সামমদি আই যাইবোগই।	আবার ইকান গইরজি বাই কুননননীর খরমইদ্যে দ্বি ি দেখির দি তারফরে সেম। ইল নাম্বার নও লহরে ফরে। খ আইচ্ছে। তকুকি আঁর গুরগুম, ওয়লার দে নাছিয়্যের লইস্যে চুইত্তেএরে সামমইদ্দে আই যাইব গই।	Test	1.16	0.661971
test_chittagong (102).wav	rec_31_audio_45.wav	Chittagong	ন? অয়। তোর দাদা নাই? আছে তো। অফিসত ন-যায়? গিয়ে। আইচছা অইয়ে তইলে। অই অড়ি! হ।	নহ। আই। দুব দারা নাই? আছে তো অফিসত ন যাই? গিয়ে আইচছা অইয়ি তইলে। ওয় ওডি খ।	Test	0.75	0.346666
test_chittagong (103).wav	rec_31_audio_46.wav	Chittagong	আরে তো মা ডাহে। তরে? অয়। আইচছা ঠিক আছে, রাহি তইলে। রাখ তইলে। বাই বাই, টা টা। ভালা তাহিস। তই যাইলে, হতা অইবো তইলে কেন? আইচছা। টাটা।	আরে তো মা ডাহ। তোরে? অ আইচ্ছা টিক আছে? রাহি ত ইলে।বাঅলে বাই বাই টাটাই। ভাড়া তাহি? চ্যাটা। তুই ভাডলা তাহিস। আইখন? আইসছা আইসছা।	Test	0.8148148148	0.4
test_chittagong (105).wav	rec_32_audio_0.wav	Chittagong	আসসালামু আলাইকুম। ওয়া আলাইকুম সালাম। কেন আছো? আলহামদুলিল্লাহ, তুই কেন আছস? আলহামদুলিল্লাহ গম আছি। তুর আম্মুতারা বেকগুন কেন আছে?	আসালা মওয়ালিইকুম। ওয়ালাইকুম তা না? কেন আছ? আলহামদুলিল্লাত তি ক্যান আছস? বাহম লে গমাছি। তুর আম্মু তইরে বেয়কগুন ক্যান আছে।	Test	0.8947368421	0.362204
test_chittagong (106).wav	rec_32_audio_1.wav	Chittagong	আলহামদুলিল্লাহ গম আছে। আইচছা, আই তুরে হই, ওইদিন অইলদি, ফুয়া ইবা হতা নহয়নে, তো অইয়ে কি, এন্ডে অইলদি <> আফুর হতা যেততে আর আম্মু তুইললি! তো মানুষ ইবা সান্তা আফুরে এদিন সাইতো চাইল, আই হর আইনতো গেইলাম দে।	লেগবছে। হ্যা? আইসে আরতুরে খই। ওই ওইদিন অইলদি ফুয়াইবের খতা ন খয় না। তই কি? অ। এনড অইলদি চান্দপুর খত তে। তো মানুষ ইবে সাঞ্জ আফরদি সাইতু সাইয়লে আঁই খর ইনতাম গেলাম দে	Test	0.9210526316	0.512562
test_chittagong (107).wav	rec_32_audio_2.wav	Chittagong	কিন্তু হর আনিদদে এদিন আরতুন এক ঘন্টা রাতিয়া বঅই তাহন ফইরগে, হর ইবারলাই, কারেন্ট নআচছিলদে, এগিন ইস্ত্রি গরন ফরিবো। তো আই এক ঘন্টা বই থাইকিলাম দে, এই সময়ত হদদে, অনে মেফুয়ার ছবি দেহাইতো, তই আর মোবাইলত চার্জ নআচছিলদে,আই ঘরত রাহি গেইলাম।	কিন্তু র খরাইরদি না তুন এক ঘন্টা অইলদে রাতিয়ে বই থাহম ফরগি। তুই খরিবার লাই কারন ন আছি দি নিষটির গরম ফরিবু। তো আঁই এক ঘন্টা বই থাইক্কিলাম দে হই সময়ত হরদে অনে মেয়েফরো ছবিদে হাইতার যেলা ম ওবাইলে চাচন আছেদ, আগার রাখি গেইয়লাম	Test	0.8780487805	0.424892
			ইন লই অডি. বউত হতা হইয়ে আর আম্মারে. কি বলে. ব্যাংকার	য়িান লইয়ার অইলদি বউত খতা খই. আঁর আমমলে কী বলে			

Figure 6.5: Model inferences samples on Chittagong data

A	B	C	D	E	F	G	H
file_name	External_ID	Region	Transcriptions	Predictions	Split	WER	CER
test_narsingdi (1).wav	rec_23_audio_46.wav	Narsingdi	তারা তো ওই খারাপ খেইলাই আসে না। তারা ভালো, ওই লিগগুলাতে ভালো খেলে, ডোমিস্টিকে ভালো খেলে। পরেই তো ওদেরকে জাতীয় দলে ডাকা হয়। হ। কিন্তু, কী করবি কে? প্লেয়ার তো আর একটা দুইটা না। প্লেয়ার তো অনেক।	তা তো ওই খারাপ খেলারর সে না, তারা ভাল। ওই লিগগুলাতে ভালো খেলে, ডমেস্টিকে ভালো খেলে। পরে তো অদেরকে রাতিয় দরে ঢাকা হয়। হ কিন্তু কী করবু ক! প্লেয়ার তো আর একটা দুইটার না প্লেয়ার তো অনেক।	Test	0.5405405405	0.1631578947
test_narsingdi (10).wav	rec_23_audio_55.wav	Narsingdi	এডা বলতেই হবে। হ, এডা আমিও মানি। বাংলাদেশ উন্নতি করতেছে, আরও উন্নতি করবে। নতুন, নতুন প্লেয়ার আসবে। <> তবে, আমাদের যে পঞ্চপান্ডব ছিলো, এরকম পাওয়া যাইবো না রে আর।	এডা বলতিই আমারকে।। হ, এডা আমিও মার। মার শুন্নতি করতেছে, আরও উন্নতি করবে। নতুন নতুন প্লেয়ার হাসবে। ই তবে আমাদের ওইযে পঞ্চ পঞ্চটব ছিল, এরকম পাওয়া যাবো নারে আর।	Test	0.6206896552	0.21875
test_narsingdi (100).wav	rec_24_audio_66.wav	Narsingdi	এডারে কী করমু? খাচ্চর। যেডাইওক, বাদ দে। <> ফাচলে ছুয়া। ইদের ফরে তো ঘুরতে যাওয়ার যায়গা এক নানু বাড়িওই, না? নাহ। আর? আর কই? ইদে সালামি নিবা না কি আমাকে? সালামি।	ডাই কি করম খাবচল? ওই যেইডাই, বাদ্যে ম। দের ফরে তো সসচ যবদি বাইকাছু বমাটে ে না? নাহ। আঁর। আর কই। ঈদে সালামিনিবার নাকি আমা কে? সালামিমি।	Test	0.8387096774	0.417721519
test_narsingdi (101).wav	rec_24_audio_67.wav	Narsingdi	সালামি, সালামি। কেউ সালামি দে না, বাই। ওই ছাচু যে দিতো ইদের মধ্যে? দিআইছে দিছে তো। হেডা তো শপিংয়ের টাকা দিয়া দিছে। প্রায় দুই হাজার টেহা দিছে। সিরিয়াস! একটা টেহাও আমার আতে আইতো না, বইন। মা। মারে। ছাচু আমারে দিয়া দিছে। আমি সুন্দরভাবে ভদ্রভাবে মার আতে দিয়া লাইছি। হেদিন না আবার ডাহাত আইছে বইত। চুপ। আস্তে ক।	সালামি সালামি, কেউ সালামি দে না ভাই।ই ছাচুদর যদি ধডার মদ্যে দিস দিছে ত। হেডা ত সবিংয়ের টাকার দারার। হশ্রে দুই হাজার টেহা দিছে। শসে সিরিয়াস একটা টেহাও আমারা তোবই তো না ে। আমার ছাসায় আমারে দে দিছে, আমি সুন্দর বডা বদ্রে ভ্যা মারা তভিয় লাইছে। হেইদিনে আবার তায় দেছে ভাই। আচ চু, আসতে কও।	Test	0.8070175439	0.3464052288
test_narsingdi (102).wav	rec_24_audio_68.wav	Narsingdi	কিন্তু একটা ফ্যাক্ট কি, দেরি কইরা দিলে টেহাডি অক্করে রাইকা দিলোঅইলে। দিল জারা, জারা <> অবশ্য আমি আবার ইয়া এ আশুলিয়া যামু সময় দাদু এমনেও পাঁচশো টাকা দিয়া দিবো এক্সট্রা। মানে এহন ফরতেকবার আইলে এডা লাভ অইতাছে, পাঁচশো টাকা কইরা দিয়া দেয়। খাড়া। ওয়েট, ওয়েট। একটা জিনিস।	কিন্তু একটা টাক কি। দিনবরে দিলে টিয়াডিয়া এক্কাডে রাইক্কা দি অইনু। অ অছোনু কিেও অবে ি আবার ইয়া আসে যাষন সময় দ দি দে পারশো টাকা দিয় দিবো এসট্মন ুনে এহন ফরটে বাইলে এা লাউ টা, দয়ে খারা অই পয় যা	Test	0.8510638298	0.5133079848
test_narsingdi (103).wav	rec_24_audio_69.wav	Narsingdi	উল্টা-ফাল্টা কতা ক। আল্লাহু আকবর। এ নুবা? হু। ফালুদা রান্দা শিখছছ? হ। তুই কী কী রানতারছ? আমি? হুম। <> গরম পানি করতারি। আর? <> ।	অল্লাহযেকর। এন্যে বা? আলুদরান্দা শিখচস। ইতু কি কি রানতারস? আমি। হুম। আলু। গরম পানি করতারবু। হর	Test	0.88	0.484375
test_narsingdi (104).wav	rec_24_audio_70.wav	Narsingdi	গ্যাস পানি সিদ্ধ। ইয়ার কি চার্জ শেষ? না। ফালুদা না বানাইছিলি হেই দিন? কেমনে, কেমনে বানাইছছ রে? বইন, বইন, বিশ্বাস কর, এডা জাস্ট একটা কি কইতাম! মানে, আমি আমার মতো কইরা বানাইছি, ঠিক আছে? হু, হু। গুড়া দুধের মধ্যে কফি ঢাইলা, ব্লেন্ড কইরা আগে কফি বানা।	জসফানি সিদ্দো। ইয়ের কি চার শেষ? ন? খালুদনার বানাইছিলি এইদিন কেন্নে এইমনে বান ইছিলে বইন বইন। বিশ্যাস করে, এডার জাস্টট্রা কি কইতাম ডে, আমি আমার মত কইরে বানাইছি। ঠিক আছে। হ হি। গুডরা দুজের মদ্যে কফিটাইল্লা য্াি কার গে কফি মা।	Test	0.8260869565	0.2926829268
test_narsingdi (105).wav	rec_24_audio_71.wav	Narsingdi	কফি বানাইছি। নরমাল কফি, কোল্ড কফি এডার মধ্যে সিক্ক নুডুলস <> ছাইড়া দিছি ডেলি-টেলি। সিম্পল। ডেলি টে জা ডেলি কইত্তে আনছিলি ইতা? লাইক জাস্ট ডেলি। জাস্ট ডেলি আর ওই হেইডা কী আছিলো? কাপের মতো যে দেখলাম। ওইদে ফাস্টের পুডিংডি আছে না, বাচ্চারা খায়? ম্যাংগু পুডিং।	কপিবার নাইছি, রুমক কেডো ই সিন্দ নরুডু স্র য়াচাইর ে জিলিডাস। জলিইদ্বা জ্ি্ইিখ টেন শেলিন্ডা। জ্াশ জেলিস। জাশ ট জ্ইলি আরওই খেড কী আছিলো? ককেন মঅফ পরির দাছে ফাস টার পুটিং ছোটা বাচ্চানা খায়। ম্যাংগু বন।	Test	0.9333333333	0.5196850394
			বাবায় খুব। না, মজা আছে। <> তুই কিছু ট্রাই করলে আবার বাবায় খুব খুশি অয়। তো কেত্তাম এলা? লাইক। আর আমি। এইযে, ওয়েট। কোকোমেলন নার্সারি রিদমস। আর ইয়া ওই আমি কিছু ট্রাই করলে, দাদু খুশি অইতো	বাদা তা মজা শেঠা আমি কিছু ট্রাই করলে। আবা বাবা খুব খুশি অয়া, তে ডেদেলা লাই।র এইযে, ব.এড কউকবেরানের সেরিরে দেম। আর ইয়া,			

Figure 6.6: Model inferences samples on Narsingdi data

6.3 Benchmarking Performances

The benchmarking performance of different models and our fine-tuned model on our developed regional speech corpus is shown in Table 6.2 and is visualized later in Figures 6.7, 6.8 and 6.9.

Table 6.2: Benchmarking Performance

	Chittagong		Kishoreganj		Narsingdi		Narail		Rangpur		Average	
	WER	CER	WER	CER	WER	CER	WER	CER	WER	CER	WER	CER
Google ASR	0.984	0.943	0.948	0.893	0.902	0.807	0.882	0.774	0.939	0.849	0.931	0.853
Wav2Vec2 Large	0.989	0.671	1.032	0.754	0.958	0.575	0.952	0.574	0.967	0.646	0.979	0.644
Hishab Conformer	0.977	0.666	1.244	0.768	0.842	0.539	0.864	0.512	0.918	0.567	0.969	0.610
Yellowking	0.983	0.803	0.959	0.801	0.932	0.720	0.903	0.652	0.937	0.712	0.943	0.738
Tugstugi	0.942	0.633	1.058	0.887	0.791	0.476	0.758	0.453	0.830	0.549	0.876	0.599
PX5 (Proposed)	0.886	0.471	0.970	0.557	0.798	0.399	0.757	0 .364	0.866	0.466	0.855	0.451

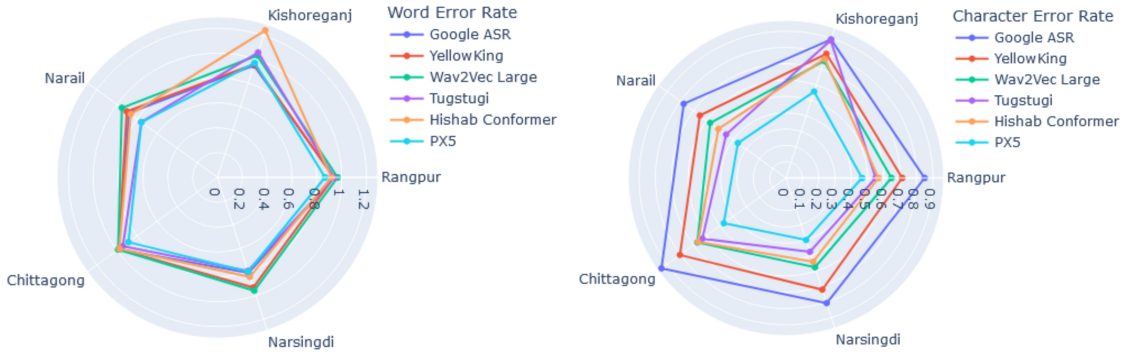

Figure 6.7: Radarplot of the model performances on regional speech corpus

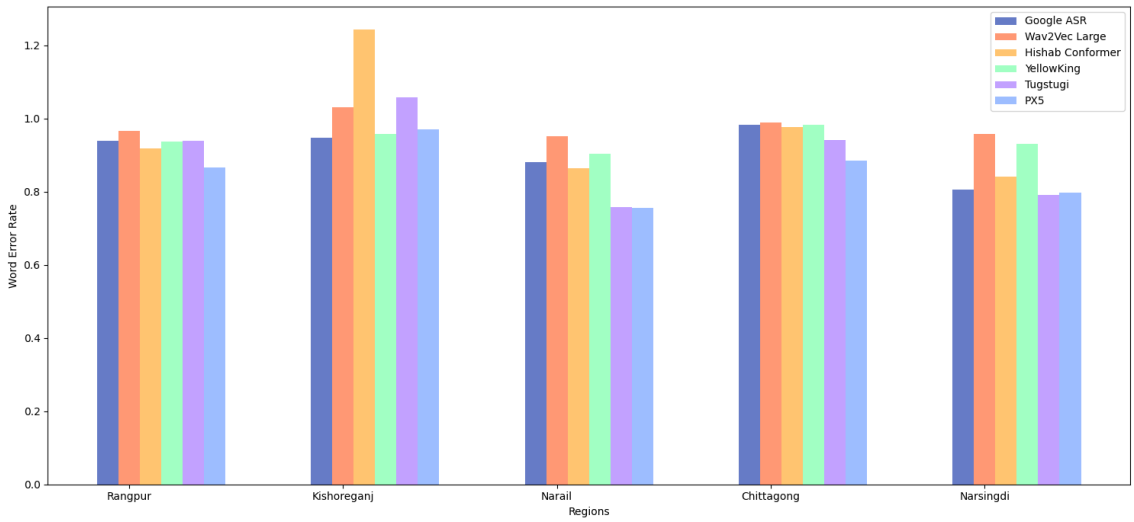

Figure 6.8: Barchart of the PX5's performance based on WER compared to the other benchmark models on regional speech corpus

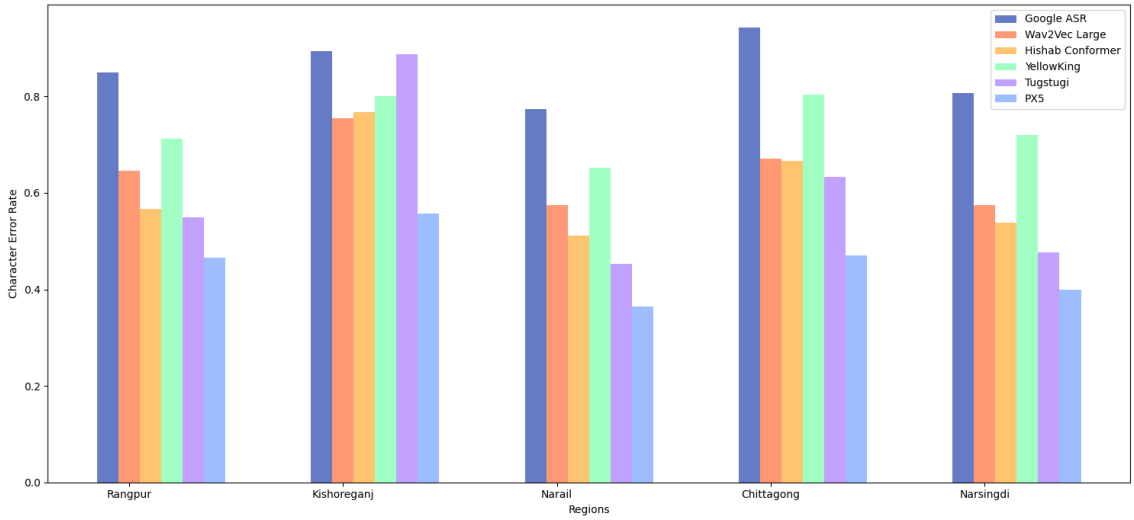

Figure 6.9: Barchart of the PX5's performance based on CER compared to the other benchmark models on regional speech corpus

6.3.1 Region-wise Benchmarking Performances

As illustrated in Figure 6.7, the proposed PX5 model outperforms all existing state-of-the-art models, demonstrating a comparatively lower word error rate and a significantly reduced character error rate. This observation is consistent with Figure 6.8 and 6.9 , which illustrates the proposed modelś performance comparing with all the other benchmark models across various regions.

Region-wise model's performances are also provided in the figures 6.10 and 6.11 for respectively Rangpur, Kishoreganj, Narail, Chittagong and Narsingdi regions.

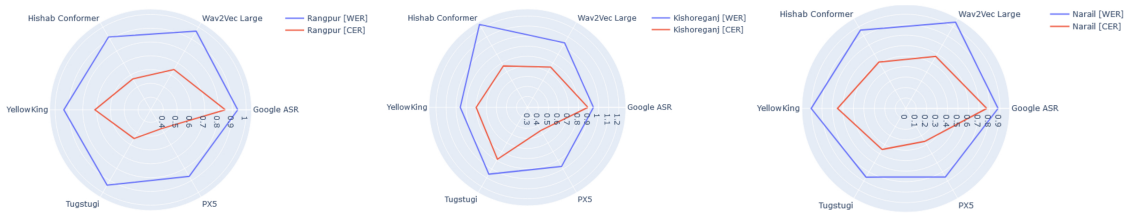

Figure 6.10: Radarplot of the model performances on Rangpur, Kishoreganj and Narail Data

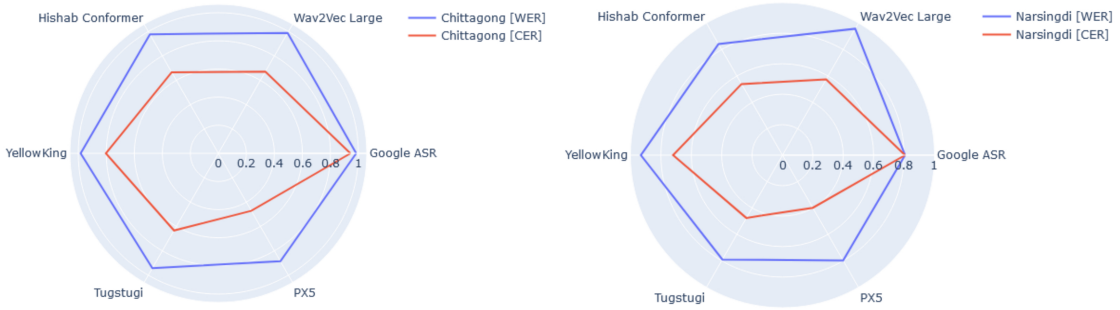

Figure 6.11: Radarplot of the model performances on Chittagong and Narsingdi Data

6.4 Inference Analysis of the Benchmark Models

Given that Bengali is a low-resource language, it has been observed that the model encounters significant difficulties in accurately transcribing named entities such as name of a person, animal, flower, fruit, or location. This issue is evident across transcription samples from different regions. Examples illustrating these challenges are provided below.

Models	Predictions	Word Error Rate (WER)	Character Error Rate (CER)
Ground Truth	হারাগাছে? হুম! তাইলে ইয়া, জমি জমাত তো আবাদ করেন নাই, না আপা? না মোর জমিন নাই, ও গরু কয়টা আছে? গরু আছে একটা। হাঁস- মুরগিও আছে না? হাঁস-মুরগিও আছে।	N/A	N/A
Google ASR	হারা করছে	1	0.939
Wav2Vec2 Large	হারা গোছেপারাক নমুতিলিয়াজুমিযবাত আবদ করেন নেইনে আপা না আমুজ জামির নাইগুরু কোয়টে আছে গরু আছে এক তাককাশমুরগি আছে না রাসমুর দিয়ে আছে।	0.828	0.432
Hishab Conformer	হারা গেছে আর নাম তাহলে জমি জমার তো আবাদ করেন নাই না আপা না আমার জমির নাই গরু কয়টা আছে গরু আছে একটা মুরগি আছে না হাস মুরগি আছে	0.69	0.264
YellowKing	কেরামত আবাদ করেন নেই না মমিনপুর কেটে গরু আছে কাশমুরের আছে।	0.793	0.655
Tutstugi	তাইলে গিয়া জমিজমা তো আবাদ করেন নাই না পা গুরু কয়টা আছে গরু আছে একটা হাঁস মুরগি আছে না হাঁস মুরগিও আছে	0.724	0.365
PX5	খারা কছেন? পরেখ না মুই। তাইলে ইয়া জমি জমার তো আবাদ করেন নাই। না আফা? না মোর জমির নাই। গরু কয়টা আছে? গরু আছে একটা। হাঁস-মুরগি আছে না? খাস-মুরগিও আছে । ও	0.621	0.209

Figure 6.12: Inference results for all the models of samples from the corpus's Rangpur segment

In Table 6.12, which displays the ground truth alongside all model predictions for a sample from the Rangpur segment of the corpus, the term হারাগাছে is either missing or inaccurately transcribed as হারা করছে or হারা গেছে. While these terms demonstrate some phonetic similarity, they are orthographically distinct and incorrect. Our model predicted খারা কছেন, which deviates significantly from the expected transcription but the overall transcription quality of our model surpasses that of other models, as evidenced by its lowest CER and WER for this sample.

Models	Predictions	Word Error Rate (WER)	Character Error Rate (CER)
Ground Truth	ব্যাঙের দেহা যায় ফাল মারে। লালা ফাড়ের মাঝে দেখবা <> ব্যাঙে ফাল ফালা। লালা ভাঙলে দেহুইনযে খালি ব্যাঙে ফালফালা। ফালফালায় কি খায়? ঘাসফড়িং। ঘাসফড়িং খায়। ঘাসফড়িং কি খায়? ঘাস খায়। আবার এই ঘাসফড়িং রে খায় কেলা? ব্যাঙে। ব্যাঙেরে খায় কেলা? সাফে। ব্যাঙেরে খায় কেলা? সাফে। সাফে। এভাবেই খাদ্য শৃঙ্খল। এভাবেই ? খাদ্য শৃঙ্খল। নিচের কোন খাদ্য শৃঙ্খল সঠিক? এইযে দেহো এইনতে বুইজ্জা ফরবা।	N/A	N/A
Google ASR	ভালো ফলাফল	1	0.976
Wav2Vec2 Large	বযাঙ্গে দেরকে ভাল মারেদেলালা ওয়ারএ মরাজা তেক বা ধানকেব ব্যাংালে ভাল ফলালালা অংলেদের নিরসিবাঙ্গে ভাল্ম হালাচিকা ঘাস করিংকা রাসকইঙ্কেইকাগাসকা আর আবার এ গাস করিংযয়োকাকেলাব্যাঙ্গে ব্যাঙ্গে একাগেলা সহবে বেঙ্গের একা গেলাসাপে এভাবে কার্শন হলেভাবে খার্যশ নিচের পুর সার্যশণ কলে জটিকেযন্যেন্দে বুড্যে হর মা।	0.968	0.566
Hishab Conformer	ভেঙ্গে হাল মারে লালাারে দেখাম ভালো খা লালা গাছ শুইং গাছ শুই গাছ আবার এই গাছ শুইং একা খেলা ব্যাঙ্গে ব্যাঙ্গে একা গেলানে একা গেলা এভাবে খাদ্য শ এভাবে খাদ্য নিচে খাদ্যল বুতে ভ	0.921	0.656
YellowKing	ভালাভাবে সিকারকে সেবা ভাবে কার্যকারিতে কোন সাত্যকিকে বুজে।	0.984	0.87
Tutstugi	ভালা ভারে আমরা দেখবা ধমকেবু ভাবে ভালা ফলা ভালা বাংলা দৈন্য ব্যাংকের ফালমার চিকা এ গাছ প্রোহিঙ্কা গাছ প্রোহিঙ্কা গাছ কা গাছ কা আবার এই গাছ প্রোহিঙ্গে একাকালা বেঙে একাকালা সাফে এভাবে খাদ্য শৃঙ্খল এভাবে খাদ্য শৃঙ্খল নিচে কুন খাদ্য শৃঙ্খল সুতিকে যেন দিবুটা ভাবা	0.889	0.561
PX5	বেঙ্গে দেহো ফাল মারে। জালা পারবা দের কইব? বো ালা কে লা ফালায। জালা ভাইঙগলে দেহুন না গা। বেযাঙগে ফাল ালেল। ফলফালা কি কা?। গাস শুড়িনকাও । গাস ফরিঙ্কে কা, গাস কা। আগা আবার এই ঘাস শুড়িঙ্গে কা কেলা? বেঙ্গে, বেঙ্গের কা কেলা? সাফেদ। বেইঙগের একা কেলা? সফে, এইভাবেই খাদ্দসি কলেভাবে খাদ্দসিগা। নিজে কুনু খাদ্দসিন কললে শোটিক এ দে য নেন বুইড্ডা পারবা।	0.937	0.477

Figure 6.13: Inference results for all the models of samples from the corpus's Kishoreganj segment

In Table 6.13, which presents the ground truth and predictions for a sample from the Kishoreganj segment of the corpus, the term ব্যাঙের is inaccurately transcribed as বযাঙ্গে or ভেঙ্গে, both of which represent distinct lexical items with differing meanings.

The model's prediction, বেঙ্গে, while orthographically divergent, exhibits phonetic similarity to the original ground truth.

Models	Predictions	Word Error Rate (WER)	Character Error Rate (CER)
Ground Truth	না আসাদ মিয়ারে দিয়া ওইছেলো কিছু। বেচা ওইছেলো। তা ওত্তে আরও বেশি উনি অচ্ছে খায়। তা ওই দখলে? দখলে না। টফি আমারে কলো যে ভাই এরকম এরকম। আমি ও নিয়ে আমার মাথা ব্যাথা নাই। আমি এর পিছনে দৈড়চ্ছি। আগে এইঠুক দেহি।	N/A	N/A
Google ASR	মাসুদ মেরে দিয়ে আইছিল কিছু বেশি বেচা আইছিল তা ওর থেকেও বেশি উনি আছে খায় তাও দখলে না উনি আমার মাথাব্যথা নাই আমি এর পিছনে ধরছি আগে ইউটিউব	0.756	0.488
Wav2Vec2 Large	নাসাদমের দিয়া ছুলু কিছু বেশ বেশায় ছিল তা ওতের বেশিও নি আছে খায়তাওই তক হলেনাট অফি আমারে কলওজ বাইয়ে রকরম রকমতামি নি আমার মাথাপেথানাই আমিদের পিছনে দরও সে আগেই টোক দি।	0.927	0.435
Hishab Conformer	দিয়া ছিল কিছু বেশি বেশি উনি আছে খায় তা দখলে না ট্রফি আমারে কল এরকমকম উনি আমার মাথা ব্যথা নাই আমি এর পিছনে দ আগেকুই	0.732	0.488
YellowKing	নাসের দিল কিছু বেসরকার্টি আমারে বাইর রনি আমার মাথা বেথানি এর পিছনে ধরে আগে দেই।	0.854	0.671
Tutstugi	নাসাদ মেয়ের দিয়ে আইস হল কিছু বেশি আইস হল তা ওটেরও বেশি উনি আসে খায় তা হই দখলে না ট্রফি আমারে কলাও যে পাই এরকম রকম তা নিয়ে আমার মাথাব্যথা নাই আমি এর পিছনে দৌড়াচ্ছি আগে এটুক দিই	0.707	0.338
PX5	নাসাদ মেয়েরে দিয়া হইছেলো কিছু। বেশি বেসা হইছেলো। তা ওতে রো বেশি উনিয়া আছে খায় তা ই াই দখলে না। টফি আমারে কলো যে ভাই এরকম রকম। তা আমি নি আমার মাথা ব্যাথা নাই। আমি এরভিছনে দৌড়চ্ছে আগে এইটুক দেহি।	0.537	0.193

Figure 6.14: Inference results for all the models of samples from the corpus's Narail segment

In a similar manner, Table 6.14, which presents the ground truth and predictions for a sample from Narail, indicates that the term আসাদ is inaccurately transcribed as নাসাদ by our model. This transcription exhibits a significant deviation from the expected output. However, it is noteworthy that the overall transcription, quality of our model exceeds that of alternative models, as demonstrated by its lowest CER and WER for this particular sample.

Models	Predictions	Word Error Rate (WER)	Character Error Rate (CER)
Ground Truth	এই ইয়া সাদিয়া আইত ফারে বলে, সাদিয়ার মা ঢাহা গেয়িল, যাইবো হইয়ে। অ, ইতারা বলে পাঁচটা বাজে উইট্যি, আসতে আসতে এগারোটা, সাড়ে এগারোটা বাজিবু এন্ডে। ইয়া তুই যাইয়ারে, ইয়ুতুতুন তোর বড় হালামণির এডে যাইবু দে? অ। ফুপিরো বাসাত? অ।	N/A	N/A
Google ASR	চোখের ভাষা	1	0.969
Wav2Vec2 Large	তি লজনথাতাদিয়ায়েতরবল সাদিয়ার মা দাহাগের ব্যবু হযীকয়ত আবার পাঁচলা বা যদতি একপেত্তায়রাতা সাহয় কঠা হবত সায কঠা বাঁসি বয়ান্ডেজেতুই জনারায়তুঃ বরাহালামনির বারা ব্যবদাও শুভির বাষা।	0.949	0.604
Hishab Conformer	সাজিয়া এটার বলে সাজিয়ার মা হই হয় বলে পাঁচটা একটা জায়গাটা জায়গায় তোল আমার নিজেরা যাব বোধহয় সুখের বাসায়	0.923	0.656
YellowKing	সাদা সাজিদাকে রাকা সানারিয়ার সুফির বা।	1	0.859
Tutstugi	সাজিয়া এ তোর বলে সাজিয়ার মা দেখা যাব হই হয়তোর পরে পাঁচটা বাজুক বান্দায় যা তুই যাহারাই তো বলা হালামনি বেড়া যাব তাহা ফকির বাসায়	0.923	0.604
PX5	ত তো ইয়া সাজিয়া ইতোর বলে সাজিয়ার মা ডাহা গেই যাবু খইয়ে। অয় তের বলে পাঁচটা বাজ উইঠতি, আইশযাইস টেগটা, সাড়া ঘটে ন আবার সাড়া ঘটে বাজিবাইনডে।য় তুই যেডা তো বলে খালামনির এডে যাইবু দে? অ ফুফ্রি বাষা। অ	0.846	0.432

Figure 6.15: Inference results for all the models of samples from the corpus's Chittagong segment

In Table 6.15, which presents a sample from Chittagong, the term সাদিয়া is inaccurately transcribed as সাজিয়া. Although the erroneous transcription bears phonetic similarity to the ground truth, it differs orthographically. The Wav2Vec2 Large model transcribed the name accurately; however, our model demonstrated superior

overall performance.

Models	Predictions	Word Error Rate (WER)	Character Error Rate (CER)
Ground Truth	আসলে রুবেলের জন্য খারাপ লাগে রে। অইও তো হিসাবে সিনিয়র প্লেয়ারই আমাদের। ওরে তেমন রেসপেক্ট দেওয়া হয় না। আর সাইফুদ্দিনের কথা বললি তো? সাইফুদ্দিন তো অইলো, মানে নিজেরে নিজে বড় মনে করলে যা হয় আর কি। এই আর কি।	N/A	N/A
Google ASR	চলো রুবেলের জন্য খারাপ লাগে রে ও তো হিসেবে দেওয়া হয় না নিজেরে নিজে বড় মনে করলে যা হয় আরকি	0.641	0.567
Wav2Vec2 Large	সল রূবের জন্য খারাপ্লাগেরহহতম হিসেবে সিনির প্রিয়ারী ম ধগরের তেমনর এক্ সপোক দেওয়া হয়নাআর্থথা ্ফ উদ্দিন্ কথাবল্লিত স্থায়ফুদ্দিন ত ল অানেদ নিজের নিজে পরমরনে গুলো যা হয় আকেএর।	0.872	0.409
Hishab Conformer	রুবেলের জন্য খারাপ লাগে হিসাবে সিনিয়রয়ার তেমন দেওয়া হয় না আর সাইফুদ্দিনের কথা বলি তোইুদিন তো নিজের নিজে বড় মনে করলে যা হয় আর কি	0.513	0.356
YellowKing	রূপের জন্য খারাপ লাগে হিসাবে সিনিয়র প্রেয়ার দে তেমন কসবা দেওয়া না সাইফউদ্দিনের কথা বলি ইদ্দিন নিজে নিজে পরনে যাওয়া হয়।	0.821	0.51
Tutstugi	আসলে রোবারদের জন্য খারাপ লাগে রে প্রয়োগ হিসাবে সিনিয়র প্লেয়ার ই আমাদের ওর তেমন এক্সপেক্ট দেওয়া হয় না আর সাইফুদ্দিনের কথা বলিত সাইফুদ্দিন তো নিজের নিজের বড় মনে করলে যা হয় আর কি	0.641	0.279
PX5	আসলে রুবের রে জনয খারাপ লাগেরে। ওইযও তো হিসাবে সিনিয়র প্লেয়ারি। ওরে। ওরে তেমন রক্সপেক্ট দেওয়া হয় না। আর সাইফুর দিনের কথা বলি তো, সাইফুর দিন তে অইলো মানে নিজেরে নিজের বড় মনে কলে যাহয় আর কি! এইআর।	0.641	0.154

Figure 6.16: Inference results for all the models of samples from the corpus's Narsingdi segment

Finally, in Table 6.16, which represents a sample from the Narsingdi segment of the corpus, the term রুবেলের is inaccurately transcribed as রুবের রে. In contrast, some English terms, such as সিনিওর and প্লেয়ার, are transcribed accurately despite their lower frequency in the corpus. This accuracy is attributable to the model's prior familiarity with these English words, which facilitates more effective learning of their orthographic forms.

In conclusion, as demonstrated in Tables 6.12, 6.13, 6.14, 6.15, and 6.16 for a single sample, our model exhibits the lowest CER and, with the exception of the Kishoreganj sample, also achieves the lowest WER compared to all other SOTA models. Overall, the performance of our model indicates superior capabilities, as corroborated by the data presented in Table 6.2.

It is also observed that our model's inference lengths demonstrate a close alignment with the ground truth while also providing reasonably accurate orthographic representations of regional dialects as shown in the above tables. We believe with more data and more training, this can be overcome.

6.4.1 Dialect Capturing Analysis

To evaluate the performance of the model in accurately capturing dialectal variations, we calculated the percentage of exact predicted words that match the ground truth dialectal transcriptions for each sample. This metric, referred to as the "Correct

Table 6.3: Dialect Detection Accuracy by District

District	Accuracy (%)
Rangpur	19.60
Kishoreganj	11.66
Narail	24.17
Chittagong	17.80
Narsingdi	17.85
Average	17.60

Dialect Percentage," provides an interpretable measure of the model's effectiveness in recognizing and predicting dialectal words accurately.

From the dataset, an average Correct Dialect Percentage of 17.6% was observed across all dialectal transcriptions. This suggests that while the model demonstrates some capability in identifying dialectal nuances, there is substantial room for improvement to ensure a more comprehensive representation of dialect-specific linguistic characteristics.

Further inspection reveals variability in performance across different samples, with some instances achieving relatively higher correct percentages, likely due to less dialectal complexity or better model alignment with the linguistic patterns of those specific dialects. Conversely, instances with lower percentages highlight challenges the model faces, such as highly nuanced or context-specific dialectal variations.

Average dialect capturing accuracy is shown in Table 6.3 and visualized in the Figure 6.17

This analysis underscores the importance of enhancing the model's dialectal recognition capabilities, which may involve augmenting the training data with more diverse and representative dialectal examples or refining the linguistic features utilized in the model.

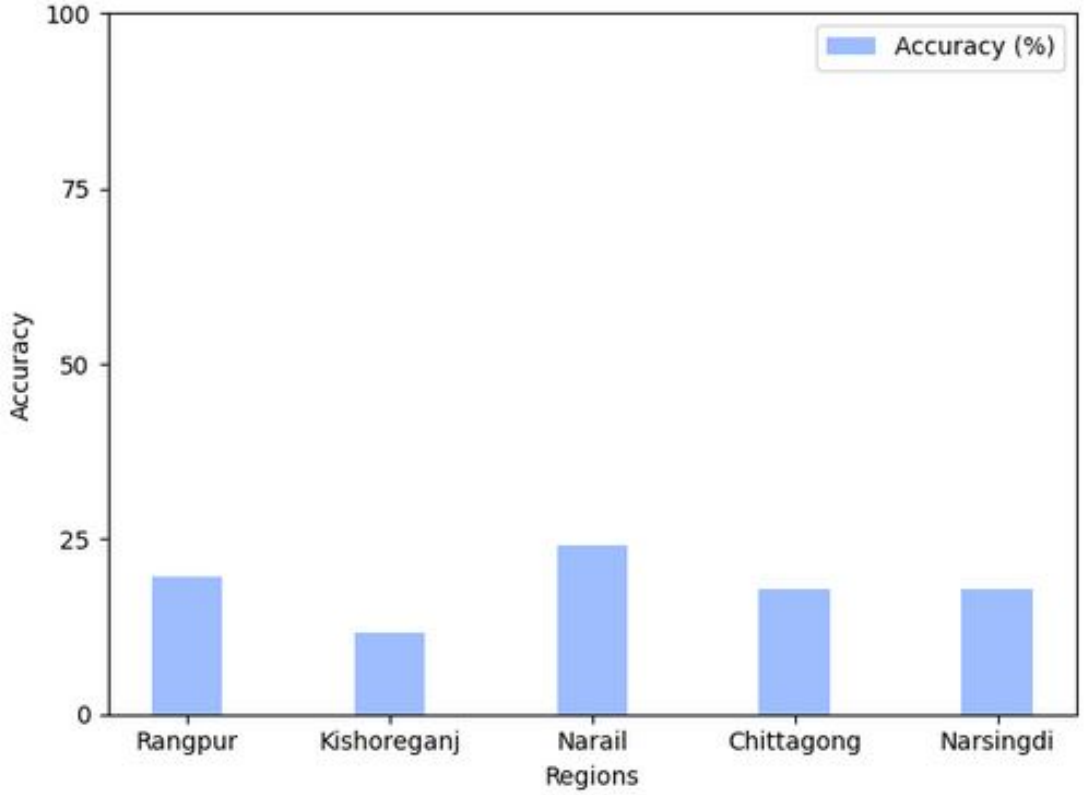

Figure 6.17: Barchart of the PX5's dialect performance of dialect-capturing accuracy

Chapter 7

Conclusion

The research presented in this paper presents an all-encompassing framework for developing the first 39 hours of regional speech corpus specifically designed to address the various regional dialects of Bangladeshi Bengali. This dataset represents the only publicly available ASR (Automatic Speech Recognition) resource specifically focused on regional dialects. We also present a detailed analysis of the linguistic complexities involved in modeling Bengali speech with regional dialects using the developed regional speech corpus. It offers significant potential for applications in federated learning, meta-learning, text-to-speech systems, and other areas requiring further exploration. The corpus, dataset, and fine-tuned model are currently undergoing refinement and enhancement. Addressing challenges such as gender bias and diversifying sentence structures within the corpus will be crucial tasks for the development of future iterations of both the dataset and the model.

We are actively collaborating with an non-profit research community named Bengali.AI, where we are working to extend this effort to include speech corpora and datasets covering regional dialects from all 64 districts of Bangladesh. The data collection, transcription and vaidation processes will be comply by the guidelines and protocols established for the initial five districts.

Furthermore, a detailed linguistic analysis of regional Bengali will be conducted to enhance the understanding of this language for linguistic researchers. Future plans include the canonization of these transcribed data to enable machine transliteration between regional accents and standard Bengali. Additionally, We intend to develop a speech-to-text model capable of converting regional text data to its standard form and vice versa.

Bibliography

[1] S. K. Chatterji, "Bengali phonetics," Bulletin of the School of Oriental Studies, University of London, vol. 2, no. 1, pp. 1–25, 1921, ISSN: 13561898. [Online]. Available: http://www.jstor.org/stable/607733 (visited on 10/20/2023).

[2] S. K. Chatterji, "Bengali phonetics," Bulletin of the School of Oriental and African Studies, vol. 2, no. 1, pp. 1–25, 1921.

[3] E. Gumbel, Statistical Theory of Extreme Values and Some Practical Applications: A Series of Lectures (Applied mathematics series). U.S. Government Printing Office, 1954. [Online]. Available: https://books.google.com.bd/books?id=SNpJAAAAMAAJ.

[4] A. Hai, Dhwonibijnan O Bangla Dhwonitottwo, 3rd. Bornomichil, 1964.

[5] M. Roy, "Some problems of english consonants for a bengali speaker of english," ELT Journal, vol. 23, no. 3, pp. 268–270, 1969.

[6] S. Hawkins and K. N. Stevens, "Acoustic and perceptual correlates of the non-nasal–nasal distinction for vowels," The Journal of the Acoustical Society of America, vol. 77, no. 4, pp. 1560–1575, 1985.

[7] B. Hayes and A. Lahiri, "Bengali intonational phonology," Natural language & linguistic theory, vol. 9, no. 1, pp. 47–96, 1991.

[8] S. Sen, Bhasar Itibritta. Ananda Publishers Private Limited, 1993.

[9] J. Owens, "Case and proto-arabic, part i," Bulletin of the School of Oriental and African Studies, vol. 61, no. 1, pp. 51–73, Feb. 1998. DOI: 10.1017/s0041977x00015755. [Online]. Available: https://doi.org/10.1017/s0041977x00015755.

[10] B. Vaux, "The laryngeal specifications of fricatives," Linguistic inquiry, vol. 29, no. 3, pp. 497–511, 1998.

[11] M. R. Ramaswamy, G. Chaljub, O. Esch, D. D. Fanning, and E. vanSonnenberg, "Continuous speech recognition in mr imaging reporting," American Journal of Roentgenology, vol. 174, no. 3, pp. 617–622, 2000, PMID: 10701598. DOI: 10.2214/ajr.174.3.1740617. eprint: https://doi.org/10.2214/ajr.174.3.1740617. [Online]. Available: https://doi.org/10.2214/ajr.174.3.1740617.

[12] Z. I. Ali, Dhanibijnaner bhumika (introduction to linguistics), 2001.

[13] N. Quoc-Cuong, P. T. N. Yen, and E. Castelli, "Shape vector characterization of vietnamese tones and application to automatic recognition," in IEEE Workshop on Automatic Speech Recognition and Understanding, 2001. ASRU '01., 2001, pp. 437–440. DOI: 10.1109/ASRU.2001.1034678.

[14] S. Bhattacharya, M. Choudhury, S. Sarkar, and A. Basu, "Inflectional morphology synthesis for bengali noun, pronoun and verb systems," in Proc. of the National Conference on Computer Processing of Bangla (NCCPB 05), Citeseer, 2005, pp. 34–43.

[15] A. S. Kaye, "Gemination in english," English Today, vol. 21, no. 2, pp. 43–55, 2005.

[16] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural 'networks," vol. 2006, Jan. 2006, pp. 369–376. DOI: 10.1145/1143844.1143891.

[17] A. Messaoudi, J. Gauvain, and L. Lamel, "Arabic broadcast news transcription using a one million word vocalized vocabulary," in 2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings, IEEE, vol. 1, 2006, pp. I–i.

[18] S. K. Mandal, B. Gupta, A. K. Datta, et al., "Word boundary detection based on suprasegmental features: A case study on bangla speech," International Journal of Speech Technology, vol. 9, no. 1, pp. 17–28, 2007.

[19] M. Wald and K. Bain, "Universal access to communication and learning: The role of automatic speech recognition," in Universal Access in the Information Society volume 6, 2008, pp. 435–447. DOI: https://doi.org/10.1007/s10209-007-0093-9. [Online]. Available: https://link.springer.com/article/10.1007/s10209-007-0093-9.

[20] B. Barman, "A contrastive analysis of english and bangla phonemics," Dhaka University Journal of Linguistics, vol. 2, no. 4, pp. 19–42, 2009.

[21] A. M. Peter Kitzing and V. L. Åhlander, "Automatic speech recognition (asr) and its use as a tool for assessment or therapy of voice, speech, and language disorders," Logopedics Phoniatrics Vocology, vol. 34, no. 2, pp. 91–96, 2009. DOI: 10.1080/14015430802657216. eprint: https://doi.org/10.1080/14015430802657216. [Online]. Available: https://doi.org/10.1080/14015430802657216.

[22] F. Alam, M. Habib, D. Sultana, and M. Khan, "Development of annotated bangla speech corpora," Sep. 2010.

[23] M. A. Berezina, D. Rudoy, and P. J. Wolfe, "Autoregressive modeling of voiced speech," in 2010 IEEE International Conference on Acoustics, Speech and Signal Processing, IEEE, 2010, pp. 5042–5045.

[24] M. M. Rashid, M. A. Hussain, and M. S. Rahman, "Text normalization and diphone preparation for bangla speech synthesis," Journal of Multimedia, vol. 5, no. 6, p. 551, 2010.

[25] J. Schalkwyk, D. Beeferman, F. Beaufays, et al., ""your word is my command": Google search by voice: A case study," in Advances in Speech Recognition: Mobile Environments, Call Centers and Clinics, A. Neustein, Ed. Boston, MA: Springer US, 2010, pp. 61–90, ISBN: 978-1-4419-5951-5. DOI: 10.1007/978-1-4419-5951-5_4. [Online]. Available: https://doi.org/10.1007/978-1-4419-5951-5_4.

[26] A. Stevenson, Oxford dictionary of English. Oxford University Press, USA, 2010.

[27] B. Das, S. Mandal, and P. Mitra, "Bengali speech corpus for continuous automatic speech recognition system," in 2011 International conference on speech database and assessments (Oriental COCOSDA), IEEE, 2011, pp. 51–55.

[28] B. Das, S. Mandal, and P. Mitra, Shruti bengali continuous asr speech corpus, 2011. [Online]. Available: https://cse.iitkgp.ac.in/~pabitra/shruti_corpus.html.

[29] H. Jégou, M. Douze, and C. Schmid, "Product quantization for nearest neighbor search," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 33, no. 1, pp. 117–128, 2011. DOI: 10.1109/TPAMI.2010.57.

[30] S. Mandal, B. Das, P. Mitra, and A. Basu, "Developing bengali speech corpus for phone recognizer using optimum text selection technique," in 2011 international conference on asian language processing, IEEE, 2011, pp. 268–271.

[31] D. Povey, A. Ghoshal, G. Boulianne, et al., "The kaldi speech recognition toolkit," IEEE 2011 Workshop on Automatic Speech Recognition and Understanding, Jan. 2011.

[32] J. I. Ali, Introduction to phonology. Maola Brothers, Dhaka, 2012.

[33] M. D. Haque, "Bhashabigganer kotha," 2014.

[34] C. J. Maddison, D. Tarlow, and T. Minka, A* sampling, 2015. arXiv: 1411.0030 [stat.CO]. [Online]. Available: https://arxiv.org/abs/1411.0030.

[35] A. K. M. Morshed, "Adhunik bhashatotto," 2015.

[36] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An asr corpus based on public domain audio books," in 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2015, pp. 5206–5210. DOI: 10.1109/ICASSP.2015.7178964.

[37] Bills, Aric, David, Anne, Dubinski, Eyal, et al., Iarpa babel bengali language pack iarpa-babel103b-v0.4b, 2016. DOI: 10.35111/5jdb-wp44. [Online]. Available: https://catalog.ldc.upenn.edu/LDC2016S08.

[38] C. A. Chapelle and E. Voss, "20 years of technology and language assessment in language learning & technology," Language Learning & Technology, vol. 20, no. 2, pp. 116–128, 2016.

[39] W. Li, S. M. Siniscalchi, N. F. Chen, and C.-H. Lee, "Improving non-native mispronunciation detection and enriching diagnostic feedback with dnn-based speech attribute modeling," in 2016 IEEE international conference on acoustics, speech and signal processing (ICASSP), IEEE, 2016, pp. 6135–6139.

[40] N. D. Londhe, M. K. Ahirwal, and P. Lodha, "Machine learning paradigms for speech recognition of an indian dialect," in 2016 International Conference on Communication and Signal Processing (ICCSP), 2016, pp. 0780–0786. DOI: 10.1109/ICCSP.2016.7754251.

[41] M. Rashid and S. Chowdhury, "Word sense ambiguity and bangla homographs a linguistic analysis," The Research Journal of Humanities, vol. 1, pp. 327–36, Jul. 2016.

[42] E. Jang, S. Gu, and B. Poole, Categorical reparameterization with gumbel-softmax, 2017. arXiv: 1611.01144 [stat.ML]. [Online]. Available: https://arxiv.org/abs/1611.01144.

[43] S. Darjaa, R. Sabo, M. Trnka, M. Rusko, and G. Múcsková, "Automatic recognition of slovak regional dialects," in 2018 World Symposium on Digital Intelligence for Systems and Machines (DISA), 2018, pp. 305–308. DOI: 10.1109/DISA.2018.8490639.

[44] ELRA, Elra-u-s 0031, 2018. [Online]. Available: http://universal.elra.%20info/product_info.php?cPath=37_39&products_id=1669.

[45] C. C. Johny and M. Jansche, "Brahmic schwa-deletion with neural classifiers: Experiments with bengali," in Proc. The 6th Intl. Workshop on Spoken Language Technologies for Under-Resourced Languages, 2018, pp. 259–263. [Online]. Available: http://dx.doi.org/10.21437/SLTU.2018-54.

[46] M. F. Khan, "Construction of large scale isolated word speech corpus in bangla," Global Journal of Computer Science and Technology, vol. 18, no. G2, pp. 21–26, 2018.

[47] M. Khan and M. Sobhan, "Creation of connected word speech corpus for bangla speech recognition systems," Asian Journal of Research in Computer Science, pp. 1–6, 2018.

[48] O. Kjartansson, S. Sarin, K. Pipatsrisawat, M. Jansche, and L. Ha, "Crowd-sourced speech corpora for javanese, sundanese, sinhala, nepali, and bangladeshi bengali," in Proc. The 6th Intl. Workshop on Spoken Language Technologies for Under-Resourced Languages (SLTU), Gurugram, India, Aug. 2018, pp. 52–55. [Online]. Available: http://dx.doi.org/10.21437/SLTU.2018-11.

[49] S. Murthy, D. Sitaram, and S. Sitaram, "Effect of TTS Generated Audio on OOV Detection and Word Error Rate in ASR for Low-resource Languages," in Proc. Interspeech 2018, 2018, pp. 1026–1030. DOI: 10.21437/Interspeech.2018-1555.

[50] TDIL, Bengali speech data – asr, 2018. [Online]. Available: http://tdil-dc.in/index.php?lang=en.

[51] J. Li, V. Lavrukhin, B. Ginsburg, et al., "Jasper: An end-to-end convolutional neural acoustic model," arXiv preprint arXiv:1904.03288, 2019.

[52] A. Nagrani, J. S. Chung, W. Xie, and A. Zisserman, "Voxceleb: Large-scale speaker verification in the wild," Computer Science and Language, 2019.

[53] R. Nordquist, Mutual intelligibility, [Online; accessed 23-October-2023], 2019. [Online]. Available: https://www.thoughtco.com/what-is-mutual-intelligibility-1691333.

[54] S. Ahmed, N. Sadeq, S. S. Shubha, M. N. Islam, M. A. Adnan, and M. Z. Islam, "Preparation of bangla speech corpus from publicly available audio & text," in Proceedings of The 12th language resources and evaluation conference, 2020, pp. 6586–6592.

[55] R. Ardila, M. Branson, K. Davis, et al., "Common voice: A massively-multilingual speech corpus," English, in Proceedings of the Twelfth Language Resources and Evaluation Conference, N. Calzolari, F. Béchet, P. Blache, et al., Eds., Marseille, France: European Language Resources Association, May 2020, pp. 4218–4222, ISBN: 979-10-95546-34-4. [Online]. Available: https://aclanthology.org/2020.lrec-1.520.

[56] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "Wav2vec 2.0: A framework for self-supervised learning of speech representations," Advances in neural information processing systems, vol. 33, pp. 12 449–12 460, 2020.

[57] N. Choudhary and D. Rao, "The ldc-il speech corpora," in 2020 23rd Conference of the Oriental COCOSDA International Committee for the Co-ordination and Standardisation of Speech Databases and Assessment Techniques (O-COCOSDA), IEEE, 2020, pp. 28–32.

[58] A. Ezzine, H. Satori, M. Hamidi, and K. Satori, "Moroccan dialect speech recognition system based on cmu sphinxtools," in 2020 International Conference on Intelligent Systems and Computer Vision (ISCV), 2020, pp. 1–5. DOI: 10.1109/ISCV49265.2020.9204250.

[59] A. Gulati, J. Qin, C.-C. Chiu, et al., "Conformer: Convolution-augmented transformer for speech recognition," 2020. arXiv: 2005.08100 [eess.AS].

[60] R. Imaizumi, R. Masumura, S. Shiota, and H. Kiya, "Dialect-aware modeling for end-to-end japanese dialect speech recognition," in 2020 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), 2020, pp. 297–301.

[61] S. Kibria, M. S. Rahman, M. R. Selim, and M. Z. Iqbal, "Acoustic analysis of the speakers' variability for regional accent-affected pronunciation in bangladeshi bangla: A study on sylheti accent," IEEE Access, vol. 8, pp. 35 200–35 221, 2020. DOI: 10.1109/ACCESS.2020.2974799.

[62] H. Liu, J. Liang, V. J. van Heuven, and W. Heeringa, "Vowels and tones as acoustic cues in chinese subregional dialect identification," Speech Communication, vol. 123, pp. 59–69, 2020, ISSN: 0167-6393. DOI: https://doi.org/10.1016/j.specom.2020.06.006. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167639320302417.

[63] I. Nigmatulina, T. Kew, and T. Samardzic, "ASR for non-standardised languages with dialectal variation: The case of Swiss German," in Proceedings of the 7th Workshop on NLP for Similar Languages, Varieties and Dialects, Barcelona, Spain (Online): International Committee on Computational Linguistics (ICCL), Dec. 2020, pp. 15–24. [Online]. Available: https://aclanthology.org/2020.vardial-1.2.

[64] S. Shivaprasad S. and M., "Identification of regional dialects of telugu language using text independent speech processing models," in International Journal of Speech Technology volume 23, 2020, pp. 251–258. DOI: https://doi.org/10.1007/s10772-020-09678-y. [Online]. Available: https://link.springer.com/article/10.1007/s10772-020-09678-y.

[65] M. H. R. Sifat, C. R. Rahman, M. Rafsan, and H. Rahman, "Synthetic error dataset generation mimicking bengali writing pattern," in 2020 IEEE Region 10 Symposium (TENSYMP), IEEE, 2020, pp. 1363–1366.

[66] R. Smith and T. Rathcke, "Dialectal phonology constrains the phonetics of prominence," Journal of Phonetics, vol. 78, p. 100 934, 2020. DOI: https://doi.org/10.1016/j.wocn.2019.100934. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0095447018300226.

[67] S. Alam, T. Reasat, A. S. Sushmit, et al., "A large multi-target dataset of common bengali handwritten graphemes," in International Conference on Document Analysis and Recognition, Springer, 2021, pp. 383–398.

[68] A. Babu, C. Wang, A. Tjandra, et al., Xls-r: Self-supervised cross-lingual speech representation learning at scale, 2021. arXiv: 2111.09296 [cs.CL]. [Online]. Available: https://arxiv.org/abs/2111.09296.

[69] V. Bhardwaj, V. Kukreja, N. Kaur, and N. Modi, "Building an asr system for indian (punjabi) language and its evaluation for malwa and majha dialect: Preliminary results," in 2021 12th International Conference on Computing Communication and Networking Technologies (ICCCNT), 2021, pp. 1–5. DOI: 10.1109/ICCCNT51525.2021.9579471.

[70] A. Bhattacharjee, T. Hasan, K. Samin, M. S. Rahman, A. Iqbal, and R. Shahriyar, "Banglabert: Combating embedding barrier for low-resource language understanding," arXiv preprint arXiv:2101.00204, 2021.

[71] J. Lee, K. Kim, and M. Chung, "Korean dialect identification based on intonation modeling," in 2021 24th Conference of the Oriental COCOSDA International Committee for the Co-ordination and Standardisation of Speech Databases and Assessment Techniques (O-COCOSDA), 2021, pp. 168–173. DOI: 10.1109/O-COCOSDA202152914.2021.9660537.

[72] M. Or Rashid, "How to process bangla linguistic-data through nlp pipeline," Dhaka University Journal of Linguistics, vol. 12, pp. 135–164, Mar. 2021.

[73] S. Sultana, M. S. Rahman, and M. Z. Iqbal, "Recent advancement in speech recognition for bangla: A survey," Int. J. Adv. Comput. Sci. Appl, vol. 12, no. 3, pp. 546–552, 2021.

[74] O. Aitoulghazi, A. Jaafari, and A. Mourhir, "Darspeech: An automatic speech recognition system for the moroccan dialect," in 2022 International Conference on Intelligent Systems and Computer Vision (ISCV), 2022, pp. 1–6. DOI: 10.1109/ISCV54655.2022.9806105.

[75] S. Alam, A. Sushmit, Z. Abdullah, et al., Bengali common voice speech dataset for automatic speech recognition, 2022. arXiv: 2206.14053 [cs.CL]. [Online]. Available: https://arxiv.org/abs/2206.14053.

[76] H. A. Alsayadi, S. Al-Hagree, F. A. Alqasemi, and A. A. Abdelhamid, "Dialectal arabic speech recognition using cnn-lstm based on end-to-end deep learning," in 2022 2nd International Conference on Emerging Smart Technologies and Applications (eSmarTA), 2022, pp. 1–8. DOI: 10.1109/eSmarTA56775.2022.9935427.

[77] K. S. Bhogale, A. Raman, T. Javed, et al., "Effectiveness of mining audio and text pairs from public data for improving asr systems for low-resource languages," arXiv preprint arXiv:2208.12666, 2022.

[78] S. Kibria, A. M. Samin, M. H. Kobir, M. S. Rahman, M. R. Selim, and M. Z. Iqbal, "Bangladeshi bangla speech corpus for automatic speech recognition research," Speech Communication, vol. 136, pp. 84–97, 2022.

[79] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," arXiv preprint arXiv:2212.04356, 2022.

[80] H. Shahgir, K. S. Sayeed, and T. A. Zaman, "Applying wav2vec2 for speech recognition on bengali common voices dataset," arXiv preprint arXiv:2209.06581, 2022.

[81] P. Swetha and J. Srilatha, "Applications of speech recognition in the agriculture sector: A review," ECS Transactions, vol. 107, no. 1, p. 19 377, 2022.

[82] M. R. I. Tomal, T. Kader, A. K. M. Masum, and M. K. A. Chy, "Bangla language dialect classification using machine learning," in 2022 4th International Conference on Electrical, Computer & Telecommunication Engineering (ICECTE), 2022, pp. 1–4. DOI: 10.1109/ICECTE57896.2022.10114552.

[83] M. Wan, J. Ren, M. Ma, Z. Li, R. Cao, and Q. Gao, "Deep neural network based chinese dialect classification," in 2021 Ninth International Conference on Advanced Cloud and Big Data (CBD), 2022, pp. 207–212. DOI: 10.1109/CBD54617.2021.00043.

[84] M. Zhai, L. Dong, Y. Qin, and F. Yu, "The research of chain model based on cnn-tdnnf in yulin dialect speech recognition," in 2022 7th International Conference on Image, Vision and Computing (ICIVC), 2022, pp. 883–888. DOI: 10.1109/ICIVC55077.2022.9886397.

[85] M. Alrehaili, T. Alasmari, and A. Aoalshutayri, "Arabic speech dialect classification using deep learning," in 2023 1st International Conference on Advanced Innovations in Smart Cities (ICAISC), 2023, pp. 1–5. DOI: 10.1109/ICAISC56366.2023.10085647.

[86] R. N. Nandi, M. H. Menon, T. A. Muntasir, et al., "Pseudo-labeling for domain-agnostic bangla automatic speech recognition," arXiv preprint arXiv:2311.03196, 2023.

[87] F. R. Rakib, S. S. Dip, S. Alam, et al., "Ood-speech: A large bengali speech recognition dataset for out-of-distribution benchmarking," Proc. Interspeech 2023, 2023.

[88] Bengali.AI, Tugstugi$_b engaliai-asr_w hisper-medium(revisionda605cc)$, 2024. DOI: 10.57967/hf/2435. [Online]. Available: https://huggingface.co/bengaliAI/tugstugi_bengaliai-asr_whisper-medium.

[89] K. Fatema, F. D. Haider, N. F. Turpa, et al., Ipa transcription of bengali texts, 2024. arXiv: 2403.20084 [cs.CL]. [Online]. Available: https://arxiv.org/abs/2403.20084.

[90] Aging voice, Accessed: 2023-09-19. [Online]. Available: https://utswmed.org/conditions-treatments/aging-voice/.

[91] Auditi Das, Wav2vec2-large-xlsr-53-bangla-common$_voice$. [Online]. Available: https://huggingface.co/auditi41/wav2vec2-large-xlsr-53-Bangla-Common_Voice.

[92] Can geography shape the way we speak? Accessed: 2023-09-19. [Online]. Available: https://blogs.scientificamerican.com/anthropology-in-practice/can-geography-shape-the-way-we-speak/.

[93] P. R. Karmaker, "Dialectical and linguistic variations of bangla sounds: Phonemic analysis,"

[94] Labelbox, Accessed: 2024-02-10. [Online]. Available: https://labelbox.com/.