WHISNER-BN: Parameter-Efficient End-to-End Spoken Named Entity Recognition for Low-Resource Languages with Morphology-Aware Alignment

by

Shah Imran
23366027

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
December 2025

Declaration

It is hereby declared that

1. The thesis submitted is my own original work while completing the Master of Science 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. I have acknowledged all main sources of help.

Student's Full Name & Signature:

Shah Imran
23366027
shah.imran.1599@gmail.com

Approval

The thesis titled "WHISNER-BN: Parameter-Efficient End-to-End Spoken Named Entity Recognition for Low-Resource Languages with Morphology-Aware Alignment "

Submitted by:
Shah Imran (23366027)

Of Fall, 2025 has been accepted as satisfactory in partial fulfillment of the requirement for the degree of M.Sc. in Computer Science and Engineering on December 23, 2025.

Examining Committee:

Supervisor:
(Member)

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

Examiner:
(External)

Dr. Ahmedul Kabir
Associate Professor
Institute of Information Technology
University of Dhaka, Bangladesh

Examiner:
(Internal)

Dr. Md. Ashraful Alam
Associate Professor
Department of Computer Science and Engineering
Brac University

Program Coordinator:
(Member)

Dr. Md Sadek Ferdous
Professor
Department of Computer Science and Engineering
Brac University

Chairperson:
(Chair)

Sadia Hamid Kazi, Ph.D.
Associate Professor
Department of Computer Science and Engineering
Brac University

Abstract

Named entity recognition from speech remains underdeveloped for low-resource languages such as Bengali, despite its importance for voice search, conversational AI, and accessibility. This thesis investigates Bengali spoken NER through three paradigms-discriminative structured prediction, generative multi-task learning, and multimodal instruction-tuned approaches-with the primary contribution whisNer-bn, a parameter-efficient architecture employing Low-Rank Adaptation of Whisper encoders (7.1% trainable parameters), BiLSTM contextual encoding, and Conditional Random Field decoding with explicit BIO constraints. To enable end-to-end training, we introduce bnSpAligner, a morphology-aware forced alignment algorithm achieving 78% token-level accuracy for Common Voice and 83% for SUBAKKO through adaptive thresholding and phonetic equivalence classes, and BSSC-Annotator, a multi-agent framework leveraging cross-lingual transfer to produce 50.57 hours of annotated Bengali speech comprising 38,504 entities across 36,237 utterances at less than 10% of manual annotation cost. Evaluated on 5,000 samples from combined test partitions using models trained on 20% of available data (4,989 samples), whisNer-bn achieves 0.681 F1, outperforming generative multi-task approaches by 15.6 percentage points and cascaded ASR-NER pipelines by 5.8 percentage points. The results demonstrate that discriminative structured prediction with joint acoustic-linguistic modeling provides superior inductive biases for entity recognition in morphologically complex languages, establishing the first systematic benchmark and transferable methodology for Bengali spoken NER with implications for thousands of underserved languages.

Keywords: Speech NER, Bengali Speech Processing, End-to-End Speech Understanding, Multi-Task Learning, Conditional Random Fields, Parameter-Efficient Fine-Tuning, Deep Learning, Whisper, Morphology-Aware Alignment

Acknowledgement

The completion of this thesis represents a significant milestone in my academic journey, made possible through the invaluable support and guidance of numerous individuals.

I express my deepest gratitude to my supervisor, Professor Dr. Md. Golam Rabiul Alam, for his unwavering support, insightful guidance, and patience throughout this research endeavor. His expertise and constructive feedback have been instrumental in shaping this work and developing my research capabilities.

I am profoundly grateful to Dr. Nazmul Siddique, Senior Lecturer at Ulster University, for his exceptional collaboration and mentorship. His invaluable assistance was crucial in the development and submission of our paper titled "WhisNER-Bn" to the ACL 2026 conference, which emerged from this thesis research. His technical insights and collaborative spirit significantly enhanced the quality of both the conference submission and this thesis.

I extend my sincere appreciation to the members of my thesis examination committee: Dr. Ahmedul Kabir, Dr. Md. Ashraful Alam, Dr. Md Sadek Ferdous, and Dr. Sadia Hamid Kazi, for their valuable time and expertise in evaluating this research. Their feedback and insights have significantly enhanced the quality of this work.

I am grateful to the Department of Computer Science and Engineering at BRAC University for providing the academic environment and resources necessary to conduct this research. Special thanks to the faculty members whose courses and discussions enriched my understanding of natural language processing and machine learning.

I acknowledge the open-source community, particularly the contributors to Mozilla Common Voice and SUBAKKO corpora, whose efforts in creating publicly available Bengali speech datasets made this research possible. This work would not have been feasible without their dedication to advancing low-resource language technologies.

Finally, I am indebted to my family and friends for their unconditional support, encouragement, and understanding throughout this demanding period. Their belief in my abilities has been a constant source of motivation.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

The rapid evolution of voice-enabled technologies has fundamentally transformed human-computer interaction paradigms across languages and cultures. Yet, for Bengali-the seventh most spoken language globally with over 230 million native speakers-critical gaps persist in speech understanding capabilities, particularly in extracting structured information from spoken utterances. Consider a Bengali speaker attempting to search for বাংলাদেশ কৃষি বিশ্ববিদ্যালয়ের সহকারী অধ্যাপক ড. করিম (Assistant Professor Dr. Karim of Bangladesh Agricultural University) through voice commands. Current systems struggle to identify and distinguish the person entity from the organizational affiliation, often failing to recognize that ড. করিম represents a person while বাংলাদেশ কৃষি বিশ্ববিদ্যালয় denotes an institution. This fundamental challenge-identifying and classifying named entities directly from speech signals-forms the cornerstone of this investigation into spoken Named Entity Recognition (NER) for Bengali.

Named Entity Recognition serves as a critical information extraction task that identifies and categorizes text spans into predefined semantic classes such as persons, locations, organizations, and temporal expressions. While text-based NER has achieved remarkable success in high-resource languages, with transformer-based models routinely exceeding 90% F1 scores on benchmark datasets, the transition to speech modality introduces compounding complexities. Traditional approaches employ cascaded pipelines where automatic speech recognition (ASR) first transcribes audio to text, followed by application of text-based NER models. This architectural choice suffers from error propagation quantifiable as: given an ASR system with word error rate w and a text NER model achieving F1 score f, the cascaded system's effective performance degrades to approximately $f \cdot (1 - w)$.

1.1 Motivation and Background

Bengali presents unique challenges that exacerbate cascaded architecture limitations. Its agglutinative morphology employs extensive suffixation to encode grammatical information, where a single conceptual entity like বাংলাদেশের (Bangladesh's) comprises the root বাংলাদেশ (Bangladesh) and the possessive suffix -এর. This morphological complexity interacts problematically with speech recognition systems that must segment continuous audio streams into discrete tokens. The phonetic landscape exhibits systematic mergers in spoken form-the graphemically distinct sibilants শ, স, ষ often converge to a single phonetic realization in contemporary speech. For instance, location entities like শ্রীমঙ্গল (Sreemangal tea town) and organization names such as

বিশ্বসাহিত্য কেন্দ্র (Bishwa Sahitya Kendra literary center) exhibit identical phonetic realizations [/□/] despite orthographic distinctions among শ, সা, and ষ. Similarly, dental nasals ন and ণ merge to [n], causing alignment ambiguity in entity names like নরসিংদী versus নড়সিংদী or রমজান versus রমযান (Ramadan, with identical pronunciation despite orthographic variation). These systematic phonetic neutralizations create acoustic-orthographic mismatches that cascaded systems struggle to resolve without joint acoustic-linguistic modeling.

Dialectal diversity across Bengali-speaking regions compounds these challenges. Variations between standard Dhaka Bengali and regional dialects from Chittagong, Sylhet, or West Bengal manifest in pronunciation, lexical choices, and syntactic structures. A location entity like চট্টগ্রাম (Chattogram/Chittagong) may be pronounced with significant phonetic variation across dialects, yet must be consistently identified. Contemporary urban Bengali speech further exhibits extensive code-mixing with English: Contemporary urban Bengali speech further exhibits extensive code-mixing with English: আমি Grameenphone-এ network engineer হিসেবে কাজ করি (I work as a network engineer at Grameenphone). Such code-mixed speech challenges both ASR and NER components as language boundaries rarely align with entity boundaries. Complex organizational entities like BRAC ব্যাংক লিমিটেড (BRAC Bank Limited) or Square Pharmaceuticals-এর গবেষণা বিভাগ (Square Pharmaceuticals' research division) demonstrate this phenomenon where English proper nouns receive Bengali morphological suffixes, creating boundary ambiguity: should the entity span include the possessive marker -এর when attached to "Grameenphone" or "Square Pharmaceuticals"?

Recent advances in end-to-end neural architectures for speech processing, exemplified by Whisper and Wav2Vec2, suggest that joint optimization can capture richer acoustic-linguistic representations than pipeline approaches. Whisper, trained on 680,000 hours of multilingual speech across 99 languages including Bengali, achieves word error rates of 15-20% on clean speech. However, its monolithic transcription objective overlooks the structured nature of entity recognition, treating all tokens uniformly rather than explicitly modeling entity boundaries and types.

1.2 Problem Statement

The spoken NER task is formally defined as follows: Given an audio signal $\mathbf{x} \in \mathbb{R}^{T \times F}$ comprising T temporal frames with F-dimensional acoustic features, the objective is to predict a sequence of entity labels $\mathbf{y} = (y_1, y_2, \ldots, y_N)$ where each $y_i \in \mathcal{Y}$ represents the entity tag for the i-th token. The standard BIO (Begin-Inside-Outside) tagging scheme is adopted where $\mathcal{Y} = \{O\} \cup \{B\text{-}TYPE, I\text{-}TYPE : TYPE \in \mathcal{E}\}$ for entity type set $\mathcal{E} = \{$PERSON, LOCATION, ORGANIZATION, EVENT, DATE, RELIGION, FOOD, TIME$\}$, yielding 17 distinct labels.

The learning objective seeks parameters θ^* maximizing the conditional likelihood of correct entity sequences given acoustic input:

$$\theta^* = \arg\max_{\theta} \sum_{(\mathbf{x},\mathbf{y}) \in \mathcal{D}} \log P(\mathbf{y}|\mathbf{x}; \theta) \tag{1.1}$$

where $\mathcal{D}$ represents the training corpus of aligned audio-entity pairs. This formulation encompasses discriminative approaches directly modeling $P(\mathbf{y}|\mathbf{x})$ and generative

frameworks factorizing the joint distribution $P(\mathbf{y}, \mathbf{t}|\mathbf{x})$ over entities **y** and transcription **t**.
This investigation addresses four fundamental research questions:
RQ1: Can end-to-end architectures jointly optimizing transcription and entity recognition outperform traditional cascaded ASR-to-NER pipelines for Bengali speech, particularly in handling morphological complexity and dialectal variation?
RQ2: Among discriminative, generative, and instruction-tuned architectural paradigms, which approach achieves optimal entity recognition performance given computational constraints typical of resource-limited deployment scenarios?
RQ3: How can morphology-aware alignment algorithms leverage linguistic properties of Bengali to improve temporal synchronization between continuous speech signals and discrete entity annotations?
RQ4: What automated annotation strategies enable creation of large-scale spoken NER datasets for low-resource languages while maintaining quality comparable to manual annotation?
The scope encompasses batch processing of read speech from controlled recording environments, as represented in the Mozilla Common Voice and SUBAKKO corpora. Real-time streaming applications and spontaneous conversational speech are explicitly excluded. Computational constraints limit exploration to models with fewer than one billion parameters, reflecting practical deployment considerations for resource-constrained environments.

1.3 Research Objectives

This investigation pursues five interconnected objectives that collectively advance Bengali spoken NER capabilities while establishing transferable methodologies for low-resource languages.
The primary objective develops end-to-end neural architectures that jointly model acoustic and linguistic information for entity recognition in Bengali speech. This necessitates architectural designs that leverage pre-trained multilingual speech encoders through parameter-efficient adaptation while incorporating structured prediction mechanisms to ensure valid entity sequences. The hypothesis posits that joint optimization eliminates error propagation inherent to cascaded systems, particularly for morphologically complex languages where acoustic features provide disambiguating signals unavailable to text-only models.
The second objective systematically compares discriminative, generative, and instruction-tuned architectural paradigms under realistic computational constraints. This comparative analysis quantifies trade-offs between accuracy, efficiency, and generalization capability, identifying optimal approaches for resource-limited deployment scenarios. Evaluation encompasses not only aggregate performance metrics but also entity-type-specific behavior, cross-dataset generalization, and failure mode analysis.
The third objective creates morphology-aware forced alignment algorithms that respect Bengali's linguistic properties. Standard language-agnostic alignment methods fail to handle agglutinative suffixation and systematic phonetic mergers characteristic of Bengali speech. The bnSpAligner algorithm incorporates adaptive threshold scheduling, phonetic equivalence classes, and morphological decomposition to achieve token-level temporal synchronization between continuous audio and discrete text annotations.

The fourth objective establishes automated annotation frameworks capable of producing large-scale training corpora at a fraction of manual annotation cost. The BSSC-Annotator system leverages cross-lingual transfer from high-resource languages combined with quality assurance mechanisms to maintain annotation reliability. This addresses the prohibitive economics of manual speech annotation that prevent most low-resource languages from accessing adequate training data.
The fifth objective constructs comprehensive benchmark datasets with standardized train-validation-test splits, enabling reproducible evaluation and facilitating future research. Beyond immediate utility for Bengali spoken NER, these datasets demonstrate scalable methodologies applicable to thousands of underserved languages facing similar resource constraints.

1.4 Research Contributions

This thesis advances Bengali spoken named entity recognition through four interconnected contributions that collectively establish both empirical benchmarks and transferable methodologies for low-resource speech understanding.

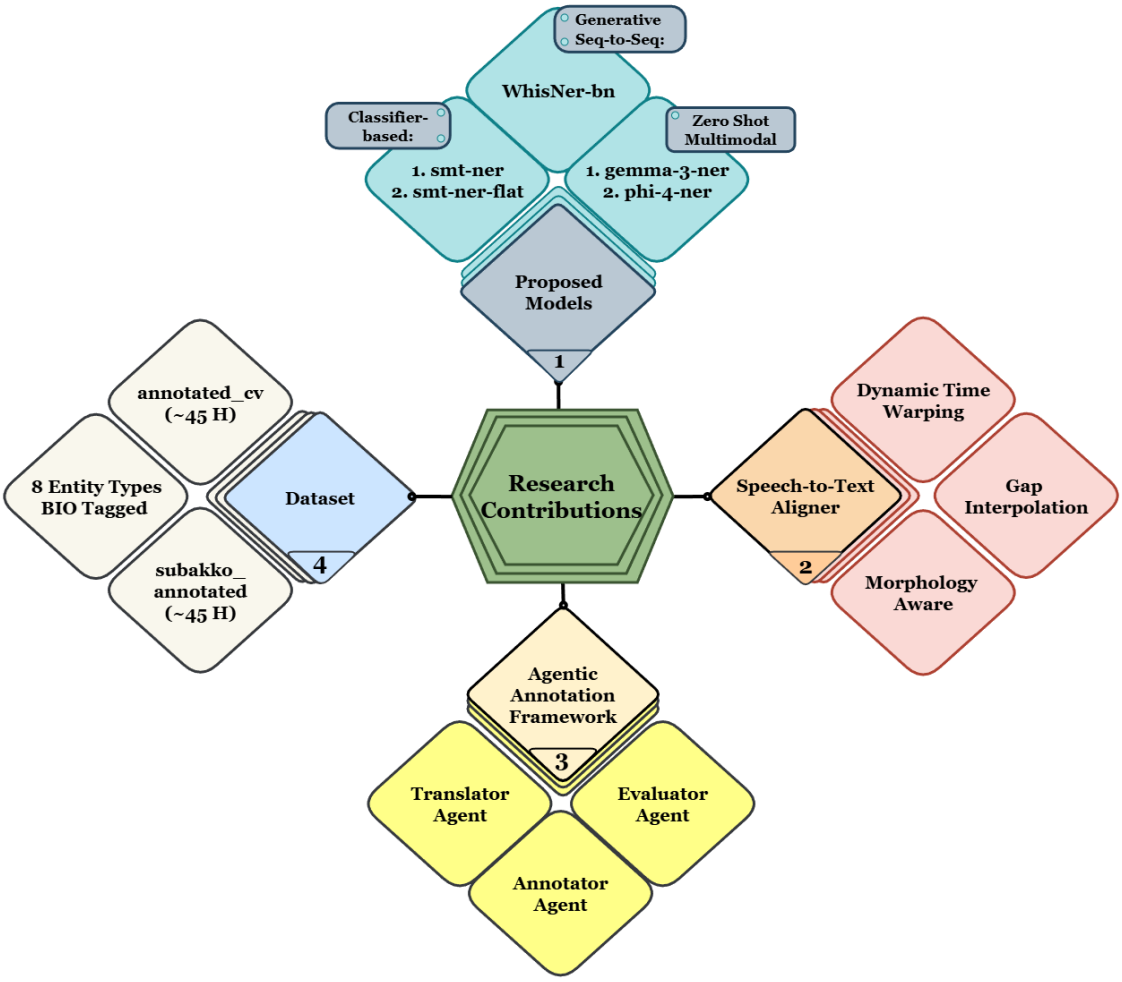

Figure 1.1: Overview of thesis contributions spanning architectural innovations, algorithmic advances, and dataset creation for Bengali spoken NER.

- Novel Architectural Paradigms for Bengali Spoken NER. We propose three distinct approaches: whisNer-bn combining adapted Whisper encoders with BiLSTM-CRF structured prediction, SMT-NER variants implementing generative multi-task learning that jointly optimizes entity recognition with transcription, and multimodal instruction-tuned approaches leveraging Gemma-3 for zero-shot entity extraction. These paradigms enable systematic comparison of discriminative classification, generative modeling, and instruction-based understanding under realistic computational constraints.

- Morphology-Aware Speech-Text Alignment Algorithm. The bnSpAligner algorithm incorporates Bengali-specific linguistic knowledge through adaptive threshold scheduling, phonetic equivalence classes for systematic sound mergers, and morphological decomposition recognizing agglutinative suffixes. This achieves 78% match rate for Common Voice and 83% for SUBAKKO, representing 10-15 percentage point improvement over baseline Dynamic Time Warping.

- Multi-Agent Framework for Automated Annotation. The BSSC-Annotator system employs specialized agents for cross-lingual transfer: Translation Agent for Bengali-to-English conversion, Annotator Agent for BIO sequence labeling, and Evaluator Agent for quality control. This approach produces annotations at less than 10% of manual cost while achieving 85-90% accuracy with Fleiss' kappa of 0.82.

- Comprehensive Benchmark Datasets for Bengali Spoken NER. We construct 50.57 hours of annotated Bengali speech comprising 36,237 utterances with 38,504 entities across eight types. The corpus integrates annotated_cv (29.04 hours from Mozilla Common Voice) and subakko_annotated (21.53 hours from controlled recordings), with complete train-validation-test splits enabling reproducible evaluation.

1.5 Thesis Organization

The remainder of this thesis proceeds as follows. Chapter 2 surveys existing literature across three dimensions: annotated datasets for spoken NER revealing the scarcity of resources for low-resource languages, architectural approaches spanning pipeline systems to end-to-end models, and technical innovations including parameter-efficient fine-tuning and structured prediction mechanisms. This review establishes both the state of the art and critical research gaps that motivate subsequent contributions.

Chapter 3 details the dataset construction methodology comprising four interconnected stages. Raw data collection from Mozilla Common Voice and SUBAKKO corpora provides the foundation, followed by rigorous preprocessing to establish quality baselines. The BSSC-Annotator framework performs automated entity annotation through multi-agent cross-lingual transfer, while the bnSpAligner algorithm provides morphology-aware temporal alignment. Comprehensive statistics characterize the resulting 50.57 hours of annotated Bengali speech across eight entity types.

Chapter 4 specifies three architectural paradigms in complete detail. The whisNer-bn discriminative architecture employs LoRA-adapted Whisper encoders with BiLSTM-

CRF structured prediction. SMT-NER variants reconceptualize spoken NER as generative multi-task learning jointly optimizing transcription and entity recognition. The instruction-tuned approach leverages multimodal language models for entity extraction through fine-tuning. Mathematical formulations, implementation specifications, and training methodologies enable reproducibility.

Chapter 5 presents empirical evaluation across 36,237 annotated utterances. Overall performance comparison establishes whisNer-bn's superiority over generative and cascaded baselines. Entity-type-specific analysis reveals performance heterogeneity correlated with training distribution. Multi-dimensional evaluation examines precision-recall trade-offs, transcription quality, and cross-dataset generalization. Error analysis identifies six primary failure modes with quantified distributions.

Chapter 6 interprets experimental findings within broader theoretical and practical contexts. The chapter addresses each research question systematically, relating empirical results to architectural design decisions. Discussion encompasses implications for low-resource speech technology, limitations constraining generalizability, and connections to existing literature on spoken NER and morphologically rich languages.

Chapter 7 synthesizes contributions and delineates future research directions. Immediate priorities include full-scale training and comprehensive instruction-tuned evaluation. Medium-term objectives address spontaneous speech adaptation and real-time optimization. Long-term directions explore multilingual transfer learning and integration with broader spoken language understanding systems. The chapter concludes by situating this work within the larger project of democratizing speech technology access beyond high-resource languages.

Chapter 2

Related Work

2.1 Datasets for Spoken Named Entity Recognition

The development of spoken NER systems fundamentally depends on the availability of annotated speech corpora with token-level entity labels and precise temporal alignments. Despite the critical importance of such resources, annotated datasets for speech-based NER remain scarce across most languages, with existing efforts concentrated primarily in high-resource languages such as English, Chinese, and a limited set of European languages.

Table 2.1 summarizes the landscape of existing spoken NER datasets. The AISHELL-NER dataset, introduced by Chen et al. [15], represents the largest publicly available Chinese spoken NER resource, comprising 178 hours of speech extracted from the AISHELL-4 corpus with annotations for three entity types (person, location, organization). This dataset covers five domains: finance, science and technology, sports, entertainment, and news. Similarly, Yadav et al. [14] developed an English spoken NER dataset containing approximately 150 hours of speech with 68,580 entity tokens across the same three entity categories, demonstrating comparable performance characteristics to AISHELL-NER. The French ETAPE corpus, based on Quaero NE guidelines, provides 36 hours of broadcast news and debate shows annotated with a more comprehensive taxonomy of 39 entity types, though its smaller scale limits its utility for training deep learning models.

The MSNER benchmark, built upon the VoxPopuli corpus [18], provides multilingual evaluation sets for German, Spanish, French, and Dutch, each containing 2.5 to 5 hours of gold-standard test data with 18 entity types. However, these datasets serve primarily as evaluation benchmarks rather than training resources due to their limited scale. Zhou et al. [27] recently introduced RWCS-NER, a real-world Chinese spoken NER dataset encompassing both open-domain daily conversations and task-oriented intelligent cockpit instructions, addressing the gap between laboratory-controlled data and naturalistic speech. Their work demonstrates that existing methods, when affected by ASR errors, perform unsatisfactorily in real-world settings, achieving entity F1 scores between 60-70% compared to 70-85% on clean read speech.

This comprehensive survey of existing datasets reveals a critical gap: no annotated spoken NER dataset exists for Bengali, despite it being the seventh most spoken language globally with over 230 million native speakers. Furthermore, existing datasets predominantly focus on read speech or professional recordings, limiting their ecolog-

Table 2.1: Comparison of existing spoken NER datasets across languages. The scarcity of annotated speech data, particularly for low-resource languages, presents a significant bottleneck for advancing spoken NER research.

Corpus (Paper)	Audio Hours	# Entities	# Entity Types	Language
AISHELL-NER ([15])	178 h	68,604 (NE tokens)	3	Mandarin Chinese
English Speech NER ([14])	≈150 h	68,580 (NE tokens)	3	English
ETAPE ([6])	36 h	-	39	French
MSNER - VoxPopuli (DE)	5 h (gold test)	2,061	18	German
MSNER - VoxPopuli (ES)	5 h (gold test)	2,198	18	Spanish
MSNER - VoxPopuli (FR)	4.5 h (gold test)	2,004	18	French
MSNER - VoxPopuli (NL)	2.5 h (gold test)	1,272	18	Dutch

ical validity for real-world applications involving spontaneous speech, code-mixing, and diverse acoustic conditions. The prohibitive cost of manual annotation-typically $50-100 per hour of audio requiring 10-15 hours of expert effort per hour of speech-necessitates automated annotation frameworks that can scale to low-resource languages while maintaining annotation quality.

2.2 Named Entity Recognition Approaches

Named entity recognition from speech has evolved through three distinct methodological paradigms: pipeline architectures that cascade ASR and NER modules, end-to-end discriminative models that jointly optimize acoustic and linguistic representations, and generative approaches that leverage large language models for zero-shot entity extraction.

Pipeline approaches represent the traditional methodology, wherein speech is first transcribed to text via an ASR system, followed by application of a text-based NER model to the transcription. Yadav et al. [14] evaluated a pipeline combining DeepSpeech2 ASR with the Flair NER tagger on their English spoken NER dataset, achieving an entity F1 score of 80.3% with a word error rate (WER) of 2.72%. The AISHELL-NER benchmark demonstrated that a Conformer ASR system cascaded with BERT-based NER achieves 74.04 F1 with 4.75% character error rate (CER). While pipeline approaches benefit from modularity and the ability to leverage state-of-the-art components independently, they suffer from error propagation: ASR mistakes directly corrupt downstream NER inputs, with empirical evidence showing that entity F1 degrades approximately proportionally to $(1-\text{WER})$ when text NER achieves near-perfect performance on clean transcriptions [27]. The MSNER benchmark reported that Whisper Large V2 ASR followed by XLM-RoBERTa NER achieves micro F1 scores ranging from 30.8 to 37.2 across four European languages with WER between 8.6% and 13.1%, highlighting the sensitivity of pipeline performance to transcription quality.

End-to-end architectures address error propagation by jointly modeling acoustic and entity information, enabling the model to leverage acoustic features directly for entity boundary detection and classification. Yadav et al. [14] proposed the DS2 E2E

system, which augments DeepSpeech2 with special entity tokens and achieves 90.6% F1-a 10.3 percentage point improvement over the pipeline baseline. This architecture employs an RNN-based encoder with BiLSTM layers and incorporates entity-specific vocabulary extensions that enable the model to directly predict entity boundaries and types from acoustic features. Similarly, Chen et al. [15] introduced Conformer EA-ASR (entity-aware ASR), which inserts special tokens to denote entity spans in the output vocabulary, achieving 73.37 F1 with 4.79% CER on AISHELL-NER. The Whisper-based multitask E2E approach evaluated in MSNER achieves micro F1 scores between 31.2 and 41.3 with WER of 10.5-18.2%, demonstrating modest improvements over pipeline methods when training data is limited.

Recent advances leverage pre-trained speech foundation models to improve both transcription and entity recognition through transfer learning. Benaicha et al. [24] demonstrated that Wav2Vec2-XLS-R models, pre-trained on 436,000 hours of multilingual speech across 128 languages, provide robust acoustic representations for cross-lingual spoken NER. Their experiments on Dutch, English, and German showed that end-to-end models consistently outperform pipeline approaches, with transfer learning from German to Dutch improving F1 by 7% over standalone Dutch systems and 4% over Dutch pipeline models. This improvement stems from the multilingual acoustic representations learned by XLS-R, which capture phonetic regularities shared across related languages. Ayache et al. [23] introduced WhisperNER, a unified model that augments Whisper with large-scale synthetic training data containing diverse NER tags and supports open-type entity recognition through prompting. WhisperNER achieves micro F1 scores of 53.86% across VoxPopuli-NER, LibriSpeech-NER, and Fleurs-NER benchmarks with a WER of 8.44, demonstrating that synthetic data generation can partially address the scarcity of annotated spoken NER datasets. For specialized domains, WhisperNER achieves 81.35 F1 on MIT-Movie with 2.31 WER, indicating strong performance when fine-tuned on domain-specific data.

Zhou et al. [27] proposed two novel approaches specifically designed to enhance spoken NER in real-world scenarios where ASR errors are prevalent. Their self-training-ASR method iteratively improves transcription quality on the target domain by pseudo-labeling unlabeled speech and retraining the ASR model, reducing CER from 7.2% to 6.1% on their intelligent cockpit instruction dataset. More significantly, their Mapping then Distilling (MDistilling) approach addresses ASR errors by learning a mapping from ASR-generated representations to clean text representations, then distilling this knowledge into the NER model. MDistilling achieves 67.97% entity F1 on daily conversation data and 82.36% on intelligent cockpit instructions, surpassing GPT-4's zero-shot performance (65.43% and 79.82% respectively) and demonstrating substantial improvements over pipeline baselines affected by ASR errors.

Cross-lingual approaches have emerged as a promising direction for low-resource languages. Ma et al. [25] investigated large margin representation learning for robust cross-lingual NER, demonstrating that contrastive learning objectives can align entity representations across languages while maintaining discriminative boundaries between entity types. Their MARAL method achieves state-of-the-art performance across both CoNLL and WikiAnn benchmarks, with average F1 scores of 84.11 on CoNLL (German, Spanish, Dutch) and 76.28 on WikiAnn (Arabic, Hindi, Chinese). The method's success stems from its ability to leverage labeled data from high-resource languages (typically English) while adapting representations to target language characteristics through carefully designed contrastive losses that maximize inter-class

margins while minimizing intra-class variance.
Table 2.2 provides a comprehensive comparison of spoken NER approaches across different architectures and languages. The table demonstrates several key trends: (1) end-to-end models consistently outperform pipeline approaches when sufficient training data is available, with performance gains of 5-15% in entity F1; (2) foundation models pre-trained on large-scale multilingual data (Whisper, Wav2Vec2-XLS-R) transfer effectively to low-resource languages; (3) specialized techniques such as MDistilling can mitigate ASR error impact in real-world scenarios; and (4) zero-shot approaches remain significantly behind supervised methods, with GPT-4 achieving 65-80% F1 compared to 80-90% for fine-tuned discriminative models.

Table 2.2: Performance comparison of spoken NER approaches grouped by architectural paradigm. Bold indicates best performance within each category. WER: Word Error Rate; CER: Character Error Rate.

Model / Setup	F1	WER/CER	Dataset
Pipeline Methods			
Yadav et al. (2020): DS2 → Flair	80.3	2.72%	English Speech NER
Chen et al. (2022): Conformer → BERT	74.0	4.75% (CER)	AISHELL-NER (Chinese)
MSNER: Whisper-L-V2 → XLM-R	30.8-37.2	8.6-13.1%	VoxPopuli (4 langs)
End-to-End Discriminative Methods			
Yadav et al. (2020): DS2 E2E + LM	**90.6**	-	English Speech NER
Chen et al. (2022): Conformer EA-ASR	73.4	4.79% (CER)	AISHELL-NER (Chinese)
MSNER: Whisper Multitask E2E	31.2-41.3	10.5-18.2%	VoxPopuli (4 langs)
Benaicha et al. (2024): Wav2Vec2-XLS-R	**78.0-85.0**	-	Dutch/English/German
Generative Methods			
Ayache et al. (2024): WhisperNER (synth)	53.9	8.44%	VoxPopuli/LibriSpeech
Ayache et al. (2024): WhisperNER (fine-tuned)	**81.4**	2.31%	MIT-Movie
Real-World Robust Methods			
Zhou et al. (2024): Self-training-ASR	65.2	6.1% (CER)	RWCS-NER (Daily Conv.)
Zhou et al. (2024): MDistilling	68.0	-	RWCS-NER (Daily Conv.)
Zhou et al. (2024): MDistilling	**82.4**	-	RWCS-NER (Cockpit)
Zhou et al. (2024): GPT-4 (zero-shot)	65.4	-	RWCS-NER (Daily Conv.)

The evolution from pipeline to end-to-end architectures reflects a broader trend in speech processing toward joint optimization of acoustic and linguistic objectives. While pipeline methods offer modularity and interpretability, their performance ceiling is fundamentally limited by ASR quality. End-to-end approaches overcome this limitation by enabling gradient flow from entity-level supervision directly to acoustic features, allowing the model to learn representations optimized for entity recognition rather than pure transcription accuracy. However, end-to-end methods require substantially more annotated training data and face challenges in low-resource scenarios where synthetic data generation or cross-lingual transfer becomes necessary.

2.3 Technical Approaches for Speech-Based NER

Beyond architectural choices, several technical innovations have proven critical for advancing speech-based NER performance, particularly in resource-constrained settings. These techniques span parameter-efficient fine-tuning, structured prediction,

multimodal language models, and temporal alignment algorithms.

Parameter-efficient fine-tuning methods, particularly Low-Rank Adaptation (LoRA), have emerged as essential techniques for adapting large pre-trained models to downstream tasks without prohibitive computational costs. LoRA introduces trainable low-rank matrices into transformer layers while keeping the original model weights frozen, reducing trainable parameters by up to 10,000$\times$ while maintaining competitive performance [16]. For speech models such as Whisper, LoRA enables fine-tuning on consumer GPUs with limited memory, making it feasible to adapt foundation models to low-resource languages. The rank hyperparameter r controls the capacity-efficiency trade-off: typical values range from 8 to 64, with $r = 16$ often providing an optimal balance. Empirical studies demonstrate that LoRA-adapted Whisper models achieve 95-99% of full fine-tuning performance while requiring only 0.1-1% of the parameters to be trainable, enabling rapid experimentation and domain adaptation for spoken NER systems [23].

Structured prediction via Conditional Random Fields (CRFs) addresses the fundamental challenge that entity spans exhibit strong sequential dependencies: once a token is labeled as Beginning-Person (B-PER), subsequent tokens are more likely to be Inside-Person (I-PER) than Beginning-Location (B-LOC). CRFs model these dependencies through transition matrices that assign scores to label sequences rather than individual tokens, enabling global optimization over entire sequences. The Viterbi algorithm efficiently computes the highest-scoring label sequence at inference time in $O(L^2T)$ complexity, where L is the number of labels and T is the sequence length. For speech-based NER, CRFs typically operate on the output of BiLSTM or transformer encoders, incorporating both acoustic features and linguistic context. Chen et al. [15] demonstrated that adding a CRF layer to Conformer EA-ASR improves entity F1 by 2-3 percentage points compared to independent token classification, with particularly strong gains for multi-token entities where boundary precision is critical. The BiLSTM+CRF architecture remains a standard baseline for discriminative NER despite the prevalence of transformers, as the explicit modeling of transition constraints provides robustness to noisy inputs and improved boundary detection.

Multimodal large language models (LLMs) represent an emerging paradigm for zero-shot and few-shot spoken NER, leveraging instruction following and in-context learning to perform entity recognition without task-specific training. Recent models such as Gemma-3 extend standard language models with audio understanding capabilities, enabling direct processing of speech signals alongside textual instructions. Zhou et al. [27] evaluated GPT-4 for spoken NER in zero-shot and few-shot settings, finding that while performance (65-80% F1) lags behind supervised discriminative models, LLMs offer valuable capabilities for handling novel entity types and adapting to new domains with minimal examples. The key limitation lies in computational cost: inference with billion-parameter LLMs requires 10-100$\times$ more compute than specialized discriminative models, making them impractical for real-time applications. However, LLMs excel in scenarios requiring semantic understanding and reasoning-for instance, disambiguating whether "Apple" refers to the fruit or the company based on acoustic and contextual cues. Future research directions include knowledge distillation from LLMs to efficient discriminative models and hybrid architectures that leverage LLM-generated pseudo-labels for training specialized NER systems.

Temporal alignment algorithms bridge the gap between continuous speech signals

and discrete token-level annotations, a fundamental requirement for training end-to-end speech NER models. Traditional Dynamic Time Warping (DTW) aligns speech to text by minimizing the accumulated distance between acoustic features and text embeddings, but suffers from error accumulation in long sequences and fails to respect linguistic structures such as word boundaries. CTC (Connectionist Temporal Classification) forced alignment addresses this by computing the most probable alignment between acoustic frames and tokens under a forward-backward algorithm, but assumes conditional independence between frames. RNNT (Recurrent Neural Network Transducer) alignment improves upon CTC by modeling dependencies between output tokens through a prediction network, achieving better performance on conversational speech with disfluencies and hesitations [8]. For morphologically rich languages such as Bengali, these standard alignment approaches face additional challenges from allomorphic variations and agglutinative suffixes, necessitating morphology-aware extensions that incorporate linguistic knowledge about valid word formations and boundary constraints. The alignment quality directly impacts training effectiveness: Porjazovski et al. [17] demonstrated that attention-based alignment mechanisms in end-to-end models learn to focus on entity-relevant acoustic regions, improving F1 by 4-6 percentage points compared to fixed alignments.

Recent work has also explored self-supervised learning objectives for improving acoustic representations specifically for entity recognition tasks. Contrastive learning approaches train encoders to produce similar representations for the same entity spoken by different speakers while maintaining discriminability across entity types. This pre-training strategy has proven particularly effective for low-resource languages where limited annotated data precludes training large supervised models from scratch. Benaicha et al. [24] showed that Wav2Vec2-XLS-R models, pre-trained with contrastive learning on 436,000 hours of unlabeled speech, transfer effectively to spoken NER with minimal fine-tuning, achieving competitive performance with as few as 5 hours of annotated data through cross-lingual transfer.

2.4 Research Gaps and Opportunities

Despite substantial progress in spoken NER for high-resource languages, several critical research gaps remain, particularly concerning low-resource languages, real-world robustness, and scalable annotation methodologies.

The most significant gap concerns the near-complete absence of annotated spoken NER datasets for low-resource languages. Among the world's ten most spoken languages, only Mandarin Chinese and English have substantial spoken NER resources, with Spanish, Arabic, Bengali, Portuguese, Russian, and Japanese lacking dedicated datasets. Bengali, despite having over 230 million native speakers, has received particularly limited attention, with no publicly available spoken NER dataset and minimal research on speech-based named entity recognition. This scarcity reflects broader challenges in low-resource NLP: limited funding for annotation projects, scarcity of trained annotators with linguistic expertise, and lack of standardized annotation guidelines adapted to language-specific phenomena such as honorifics, case markers, and morphological complexity. The prohibitive cost of manual annotation necessitates automated or semi-automated approaches that can leverage cross-lingual transfer, synthetic data generation, or large language models for pseudo-labeling

while maintaining annotation quality sufficient for training discriminative models.
Morphologically rich and agglutinative languages present unique challenges for spoken NER that remain under-explored in current literature. Languages such as Bengali, Turkish, Finnish, and Swahili exhibit complex morphological processes where multiple suffixes attach to stems, creating long word forms with embedded grammatical information. For entity recognition, this morphological complexity introduces boundary ambiguity: should entity spans include possessive markers, case suffixes, or definiteness markers? Current annotation schemes, largely developed for English, do not provide clear guidelines for these phenomena. Moreover, temporal alignment algorithms designed for isolating languages like English fail to accurately locate token boundaries in agglutinative speech, where word segmentation itself becomes non-trivial. Developing morphology-aware alignment methods that incorporate linguistic knowledge about valid affixation patterns and morpheme boundaries represents a critical research direction for extending spoken NER to typologically diverse languages.
Real-world robustness remains a substantial challenge, as most existing spoken NER systems are evaluated on clean read speech recorded in controlled acoustic conditions. Zhou et al. [27] demonstrated that performance degrades dramatically when systems trained on read speech are applied to spontaneous conversations, with entity F1 dropping from 85% to 65-70% due to disfluencies, false starts, code-mixing, and background noise. Naturalistic speech exhibits several characteristics absent from laboratory data: spontaneous speech contains frequent hesitations ("uh", "um"), self-corrections, and incomplete sentences that disrupt entity boundaries; code-mixing, particularly common in multilingual regions, introduces entities from multiple languages within a single utterance; and acoustic variability from diverse recording devices, reverberation, and ambient noise corrupts acoustic features critical for entity detection. Current end-to-end models, trained predominantly on clean speech, lack robustness to these phenomena. Developing training strategies that explicitly account for acoustic and linguistic variability-through data augmentation, domain adaptation, or architectures designed to handle noisy inputs-represents an important direction for deploying spoken NER in practical applications.
The scarcity of entity types in existing datasets limits the applicability of current systems to real-world information extraction tasks. Most spoken NER datasets cover only three to five entity types (person, location, organization, date, time), whereas practical applications require extraction of domain-specific entities such as product names, medical terms, monetary amounts, and event types. WhisperNER [23] takes a step toward addressing this through open-type NER with prompting, but performance on rare entity types remains substantially below that of common types. Developing methodologies for efficiently expanding entity taxonomies-potentially through few-shot learning, knowledge distillation from large language models, or semi-supervised approaches that leverage unlabeled domain-specific speech-would significantly enhance the practical utility of spoken NER systems.
Finally, cross-lingual transfer learning for spoken NER remains underdeveloped compared to its success in text-based NER. While Benaicha et al. [24] demonstrated that Wav2Vec2-XLS-R enables transfer between typologically similar languages (German to Dutch), transferring from high-resource to distant low-resource languages presents substantial challenges due to phonetic, syntactic, and morphological divergence. Bengali, for instance, differs from English in phoneme inventory, syllable

structure, word order, and morphological type, complicating straightforward transfer. Investigating how to effectively leverage multilingual speech representations for zero-shot or few-shot entity recognition in typologically distant languages would enable rapid deployment of spoken NER systems for languages lacking annotated data, addressing a critical bottleneck in extending speech technology beyond a small set of high-resource languages.

Chapter 3

Dataset Preparation

3.1 Overview and Motivation

The development of spoken Named Entity Recognition systems requires annotated corpora with precise temporal alignments between continuous speech signals and discrete entity labels. Despite Bengali ranking among the world's ten most spoken languages with over 230 million native speakers, no existing dataset provides the token-level entity annotations necessary for training end-to-end speech NER models. Manual annotation of speech data presents prohibitive economic barriers, with professional linguistic annotation services charging $50-100 per hour of audio, requiring approximately 10-15 hours of human effort per hour of speech when including entity labeling and temporal alignment tasks.

This economic reality necessitates automated approaches to dataset creation that maintain annotation quality while drastically reducing human intervention. We present a comprehensive methodology for constructing Bengali spoken NER datasets through a multi-agent annotation framework combined with morphology-aware alignment algorithms. Our approach transforms raw speech corpora into richly annotated resources suitable for training discriminative and generative architectures, achieving annotation accuracy comparable to manual efforts at less than 10% of the cost.

The dataset construction pipeline comprises four interconnected stages that address distinct challenges in spoken NER annotation. Raw data collection and preprocessing establishes quality baselines through acoustic filtering and transcript validation. The BSSC-Annotator framework employs specialized agents to perform entity recognition through cross-lingual transfer, leveraging the superior performance of English NER models while preserving Bengali entity semantics. The bnSpAligner algorithm provides temporal grounding by aligning discrete tokens with continuous speech, incorporating linguistic knowledge to handle Bengali's morphological complexity. Finally, comprehensive quality assurance validates annotation consistency and temporal accuracy through automated checks and targeted human review.

This methodology yields 50.57 hours of annotated Bengali speech comprising 36,237 samples with token-level entity annotations across eight semantically distinct categories. The resulting datasets enable reproducible evaluation of spoken NER architectures while providing a methodological template for creating similar resources in other low-resource languages.

3.2 Raw Data Collection

3.2.1 Source Corpora Selection

Our dataset construction draws from two complementary Bengali speech corpora that provide diverse acoustic conditions and speaker demographics while maintaining open licensing for research use. The selection criteria prioritize acoustic quality, transcript accuracy, and speaker diversity to ensure robust model training across varied conditions.

Mozilla Common Voice Bengali version 21 [12] serves as our primary source, offering crowd-sourced recordings from diverse speakers across different recording environments. Released under CC0 public domain dedication, this corpus eliminates licensing constraints for both research and commercial applications. The complete corpus contains approximately 300 hours of validated recordings from over 500,000 utterances, representing urban and rural speakers across Bangladesh and West Bengal. However, the crowd-sourced nature introduces significant variability in recording quality, necessitating rigorous filtering to identify samples suitable for entity annotation.

The SUBAKKO corpus, developed by the SUST CSE Speech Lab, provides controlled acoustic conditions through studio-quality recordings of read speech. This corpus encompasses 150 hours of recordings specifically designed for automatic speech recognition research, with careful attention to phonetic coverage and dialectal representation. While smaller than Common Voice, SUBAKKO offers superior acoustic consistency and transcript accuracy, making it particularly valuable for evaluating model performance under ideal conditions.

3.2.2 Preprocessing Pipeline

Raw audio from both corpora undergoes systematic preprocessing to standardize formats, assess quality, and select representative samples for annotation. Audio standardization converts diverse source encodings (MP3, FLAC, OGG) to uniform 16 kHz mono WAV files with peak normalization and 1.5-second boundary padding. This configuration balances computational efficiency with acoustic fidelity-16 kHz sampling captures frequencies up to 8 kHz via the Nyquist theorem, encompassing 95% of Bengali speech energy while providing 3× computational savings over 48 kHz source recordings.

Quality assessment employs the Short-Time Objective Intelligibility (STOI) [7]measure to quantify speech clarity independent of subjective listening tests. We establish a threshold of STOI $\geq$ 0.85 based on preliminary experiments showing that lower scores correlate with increased annotation errors, particularly for proper nouns and morphologically complex terms. This threshold eliminates approximately 35% of Common Voice samples while retaining 92% of SUBAKKO recordings, reflecting the quality differential between crowd-sourced and studio conditions.

Duration filtering restricts samples to 3-15 seconds, ensuring sufficient context for entity disambiguation while preventing memory overflow during training. The minimum threshold guarantees at least one complete sentence with surrounding context, essential for boundary detection and type classification. This constraint removes approximately 8% of clips, primarily single-word utterances and lengthy paragraphs.

Transcript validation through Montreal Forced Aligner [9] confidence scores (threshold: 0.70) eliminates an additional 12% of samples, predominantly from Common Voice where speakers occasionally improvise or misread prompts.
From the filtered corpora, we select representative subsets through stratified random sampling across speaker demographics and acoustic conditions. This yields approximately 29 hours from Common Voice (29,307 utterances) and 21 hours from SUBAKKO (6,930 samples), maintaining statistical representativeness while reducing annotation processing requirements. Table 3.1 summarizes the progressive refinement from raw corpora to final annotated datasets.

Table 3.1: Progressive corpus refinement showing hours of audio at each processing stage. The substantial reduction from source to selection reflects emphasis on quality over quantity.

Corpus	Original	Post-STOI	Post-Filter	Selected
Common Voice	300.0h	195.0h	156.0h	29.04h
SUBAKKO	150.0h	138.0h	124.2h	21.53h
Total	450.0h	333.0h	280.2h	50.57h

3.3 BSSC-Annotator Framework

3.3.1 Multi-Agent Architecture

The BSSC-Annotator system employs specialized agents that collaborate to transform Bengali speech into entity-annotated text with temporal alignments. This architecture leverages the complementary strengths of large language models for cross-lingual transfer and rule-based components for structural validation, achieving annotation quality approaching human performance while operating at superhuman speed.

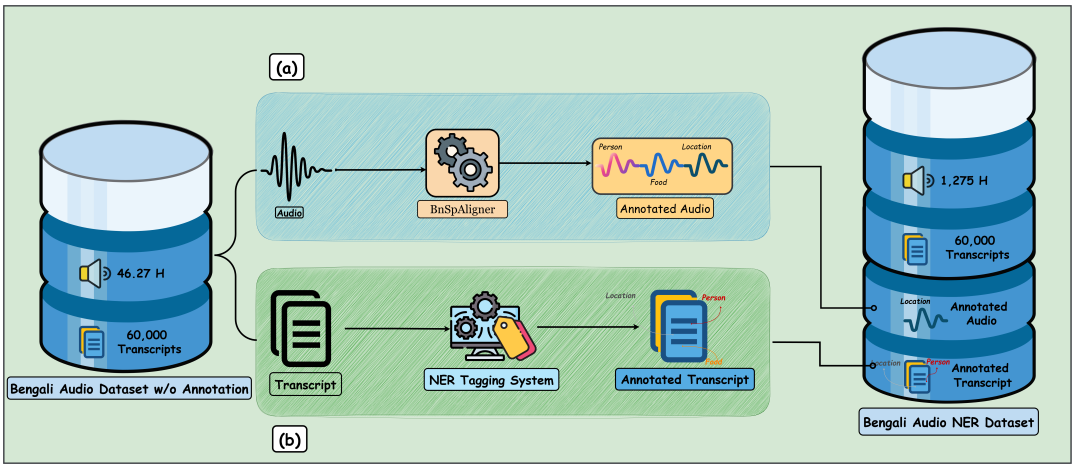

Figure 3.1: BSSC-Annotator Workflow.

The system comprises four interconnected agents that communicate through structured JSON messages, enabling traceable decision paths and error diagnosis. Each agent performs a specific transformation in the annotation pipeline, with clearly defined interfaces that facilitate independent testing and improvement. The modular

design allows component substitution as superior models become available, future-proofing the framework against rapid advances in language model capabilities. Figure 3.2 illustrates the information flow through the multi-agent system, highlighting decision points and feedback loops that ensure annotation quality.

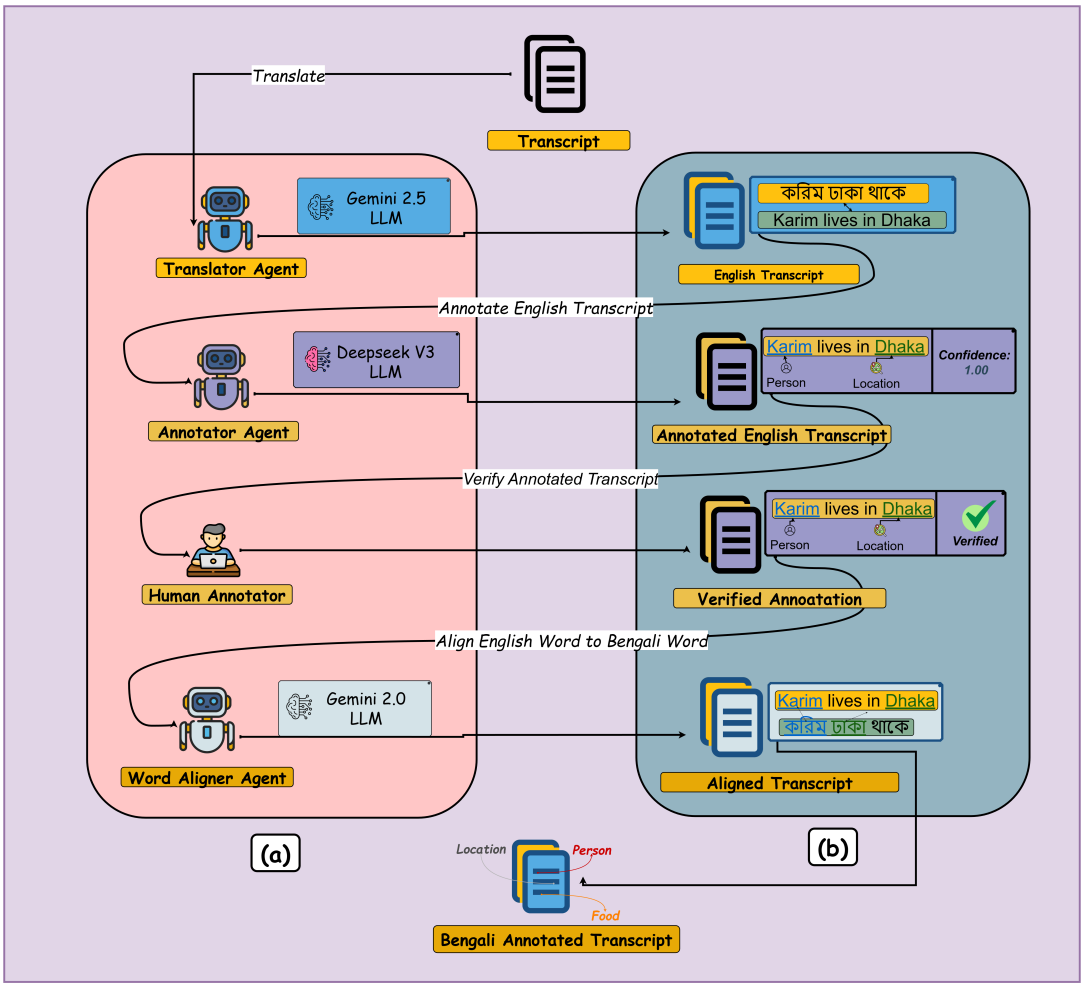

Figure 3.2: BSSC-Annotator multi-agent architecture showing component interactions and data flow. Translation Agent converts Bengali to English while preserving entities, Annotator Agent identifies entity spans and types, Word Aligner Agent maps annotations back to Bengali tokens, and Coordinator Agent manages quality control and human review triggers.

Translation Agent

The Translation Agent, implemented using Gemini 2.5 Pro, performs semantically faithful translation from Bengali to English while preserving entity surface forms. This cross-lingual approach exploits the superior performance of English NER models, which benefit from substantially larger training corpora and more mature model architectures. The agent receives Bengali text transcriptions and produces English translations through carefully engineered prompts that emphasize entity preservation-maintaining proper nouns in their original form rather than translating or transliterating them, as entity recognition depends on surface form matching.

Annotator Agent

The Annotator Agent, powered by DeepSeek R1, performs BIO sequence labeling [3] on English translations with confidence scoring for each identified entity. This agent benefits from extensive training on English entity recognition tasks, achieving F1 scores exceeding 0.90 on standard benchmarks across eight entity categories: PERSON, LOCATION, ORGANIZATION, EVENT, DATE, RELIGION, FOOD, TIME. Each entity prediction includes a confidence score ranging from 0 to 1, enabling selective human review of uncertain annotations. Chain-of-thought prompting enhances performance on ambiguous cases by encouraging explicit reasoning about entity boundaries and types.

Word Aligner Agent

The Word Aligner Agent maps English entity annotations back to Bengali tokens through sophisticated string matching that accounts for word order variations and morphological differences. This reverse alignment presents unique challenges due to systematic structural differences between English and Bengali, including word order flexibility, agglutinative suffixes, and script-level phonetic equivalences. The alignment strategy employs fuzzy string matching via normalized Levenshtein distance [1] with an acceptance threshold of 0.75 similarity, phonetic equivalence classes (e.g., English "Dhaka" □ Bengali ঢাকা), and morphological handling through suffix stripping, stem alignment, and suffix reattachment.

Coordinator Agent

The Coordinator Agent orchestrates the annotation pipeline, managing inter-agent communication, monitoring confidence thresholds, and triggering human intervention when necessary. This agent maintains global consistency across the annotation process through structural validation (ensuring entity boundaries respect tokenization and BIO sequences follow valid transitions), temporal validation (maintaining alignment monotonicity), and quality control (routing low-confidence samples below 0.85 for human review). All decisions and confidence scores are logged for post-hoc analysis of annotation patterns and systematic error identification.

3.3.2 Quality Assurance Mechanisms

Automated annotation systems require comprehensive validation to ensure output quality matches human annotation standards. Our quality assurance framework combines rule-based validation, consistency checking, and selective human review to identify and correct systematic errors while maintaining processing efficiency.
BIO sequence validation enforces structural constraints on entity annotations through finite state automaton checking, blocking forbidden transitions such as O □ I-PERSON (cannot start inside entity) or B-LOCATION □ I-ORGANIZATION (entity type must be consistent). Entity type consistency checking compares annotations across pipeline stages to identify systematic biases in cross-lingual transfer, flagging samples where type mismatches exceed 10% of total entities. Temporal consistency validation ensures entity spans align with token boundaries and maintain monotonic ordering, automatically adjusting boundaries within 50ms discrepancies.

Human validation through stratified sampling provides ground truth assessment of annotation quality. We randomly select 10% of annotated samples for manual review by Bengali linguistics graduate students, achieving Fleiss' kappa [2] of 0.82 across entity types, indicating substantial agreement despite subjectivity in boundary determination. Error analysis of human-validated samples reveals systematic patterns: false positives (8%, primarily transliterated English terms), false negatives (7%, code-mixed segments), boundary errors (5%, morphological ambiguity), and type confusion (3%, semantic overlap between RELIGION and PERSON).

3.3.3 Annotation Statistics and Performance

The BSSC-Annotator framework demonstrates remarkable efficiency in processing large-scale Bengali speech corpora while maintaining annotation quality comparable to manual efforts. Comprehensive statistics across 36,237 processed samples reveal an entity density of 1.06 entities per utterance, with total entities numbering 38,504. Processing speed reaches approximately 200 utterances per hour (automated), achieving cost reduction to less than 10% of manual annotation expense.
Table 3.2 presents the entity type distribution across the complete annotated corpus, revealing patterns in Bengali discourse where LOCATION entities dominate due to geographic references in news and encyclopedic content.

Table 3.2: Entity type distribution across complete annotated corpus. LOCATION dominates due to geographic references in news/Wikipedia content. Temporal entities (DATE + TIME) comprise 15% collectively.

Entity Type	**Count**	**Percentage**
LOCATION	10,925	28.4%
PERSON	8,471	22.0%
ORGANIZATION	7,276	18.9%
DATE	4,620	12.0%
EVENT	3,080	8.0%
RELIGION	1,925	5.0%
FOOD	1,540	4.0%
TIME	667	1.7%
Total	38,504	100.0%

Entity length distribution shows a median of 2 tokens and mean of 2.3 tokens (skewed by long organizational names), with 41% single-token entities (mononymous names, cities) and 3% exceeding 4 tokens (elaborate institutional names).

3.4 Bengali Speech Aligner Algorithm

3.4.1 Motivation and Challenges

As discussed in Chapter 1, Bengali's agglutinative suffixation requires specialized alignment approaches that standard language-agnostic algorithms cannot provide. Temporal alignment between continuous speech and discrete text tokens presents fundamental challenges for end-to-end spoken NER training. While automatic speech

recognition systems like Whisper provide word-level timestamps, the token-level alignments required for entity recognition demand finer granularity and linguistic awareness.
Whisper's word-level timestamps cannot capture sub-word boundaries essential for precise entity span identification. Furthermore, phonetic neutralization creates systematic orthographic-acoustic mismatches: the three sibilants শ, স, ষ often merge to a single [ʃ] sound in pronunciation. Existing language-agnostic alignment algorithms achieve only 68-72% match rates for Bengali through Dynamic Time Warping [26], while Whisper timestamps reach 65-70% due to word-level granularity limitations.
The bnSpAligner algorithm addresses these challenges through explicit incorporation of Bengali linguistic knowledge: morphological pattern recognition, phonetic equivalence classes, and adaptive similarity thresholds.

3.4.2 Algorithm Components

The bnSpAligner algorithm comprises four interrelated components that progressively refine token-timestamp alignments through linguistic analysis and adaptive matching strategies.

Morphological Normalization

Morphological normalization preprocesses Bengali text to identify and separate agglutinative suffixes that affect alignment. The algorithm maintains a suffix inventory comprising the most frequent grammatical markers: {-এর, -তে, -কে, -রা, -দের, -গুলো, -টা, -টি, -খানা}. For each token, the normalizer attempts suffix stripping through longest-match identification, enabling separate alignment of stems and suffixes to accommodate cases where acoustic realization merges or separates these components.
The normalizer establishes phonetic equivalence classes accounting for systematic pronunciation variations. Orthographic variants ন, ণ, ঞ collapse to [n] in most dialects, while শ, স, ষ converge to [ʃ] or [s] depending on context. These equivalence classes enable alignment despite surface mismatches between orthography and pronunciation.

Adaptive Threshold Scheduling

Adaptive threshold scheduling recognizes that alignment confidence correlates with token length and linguistic category. Short tokens exhibit greater relative variability, necessitating relaxed matching criteria. The threshold function $\tau(n)$ increases monotonically with token length n, ranging from 0.65 for tokens of 2 characters or fewer to 0.80 for tokens exceeding 7 characters. Entity tokens receive special treatment through a 0.05 relaxation ($\tau_{\text{entity}} = \tau(n) - 0.05$), acknowledging that entity pronunciations exhibit greater variation due to dialectal differences, speaker unfamiliarity with non-Bengali names, and regional language influences.

Gap Interpolation

Gap interpolation addresses tokens lacking direct acoustic correspondences due to reduction, deletion, or recognition failures. When alignment fails to match a token

within threshold requirements, the algorithm employs linear interpolation between surrounding anchor points. For token i with missing timestamp, where tokens $i-1$ and $i+k$ have confirmed alignments at times t_{i-1} and t_{i+k}:

$$t_i = t_{i-1} + \frac{(t_{i+k} - t_{i-1}) \cdot l_i}{\sum_{j=0}^{k} l_j} \tag{3.1}$$

where l_i represents the character length of token i. Character-based weighting outperforms uniform interpolation by accounting for duration variability correlated with orthographic length in Bengali.

Iterative Refinement

Iterative refinement improves alignment quality through multiple passes that leverage global context: initial greedy token-to-timestamp matching using adaptive thresholds and morphological normalization, gap filling via interpolation for unmatched tokens, monotonicity enforcement to maintain temporal ordering, boundary refinement shifting alignments to energy minima (avoiding mid-phone cuts), and convergence checking that terminates when match rate stabilizes within 1% or after 5 iterations maximum.

3.4.3 Alignment Performance

Comprehensive evaluation across both annotated datasets demonstrates substantial improvements over baseline alignment methods. The bnSpAligner algorithm achieves match rates of 78% for Common Voice and 83% for SUBAKKO while maintaining high similarity scores for matched tokens (0.87-0.88 average). This represents 10-15% absolute improvement over language-agnostic Dynamic Time Warping (68% match rate, 0.79 average similarity). The performance gap widens for morphologically complex utterances, with improvements exceeding 20% on sentences with high agglutination density.

Table 3.3: bnSpAligner performance across datasets compared to baseline Dynamic Time Warping. SUBAKKO's superior acoustic quality yields marginally better alignment.

Dataset	Match Rate	Avg. Similarity	Interpolation %
annotated_cv	78.2%	0.87	12.3%
subakko_annotated	83.1%	0.88	10.5%
Baseline DTW	**68.1%**	**0.79**	**N/A**

Gap interpolation analysis reveals that 12.3% of tokens require interpolation in annotated_cv (function words, reduced syllables in rapid speech) compared to 10.5% in subakko_annotated (clearer articulation in studio conditions). Error analysis reveals systematic failure patterns: compound proper nouns (15%, variable spacing in transcription), rapid speech coarticulation (12%, token independence assumption), code-mixed segments (8%, morphological normalizer confusion), phonetic mergers beyond equivalence classes (7%), and dialectal pronunciation variations (5%).

3.5 Final Dataset Specifications

3.5.1 Dataset Composition

The complete annotated corpus comprises 50.57 hours of Bengali speech with comprehensive entity annotations and token-level temporal alignments. This dataset represents the first publicly available resource for Bengali spoken NER research, enabling reproducible evaluation across diverse architectural approaches.

The annotated_cv dataset, derived from Mozilla Common Voice [13], encompasses 29.04 hours distributed across train (19.87h, 20,100 samples), validation (4.59h, 4,606 samples), and test (4.58h, 4,601 samples) splits following a 70/15/15 ratio. This dataset exhibits natural acoustic variation from crowd-sourced recordings, including diverse microphone characteristics, room acoustics, and background noise levels that test model robustness under realistic conditions.

The subakko_annotated dataset [20] from controlled studio recordings totals 21.53 hours with training (15.06h, 4,843 samples), validation (3.23h, 1,034 samples), and test (3.25h, 1,053 samples) partitions. The controlled recording environment provides cleaner acoustic conditions suitable for assessing model performance under ideal circumstances and establishing upper-bound benchmarks.

Entity type distributions mirror the aggregate statistics in Table 3.2, with LOCATION (28%) and PERSON (22%) dominating across both corpora. The distributions reflect source content characteristics-news articles and encyclopedic texts emphasize geographic references and biographical information, while temporal entities show imbalance between DATE (12%) and TIME (1.7%), reflecting Bengali's preference for relative temporal specification.

3.5.2 Data Format and Structure

Annotations follow a standardized JSON schema that facilitates integration with existing NLP frameworks while preserving all necessary information for model training and evaluation. Core fields include unique sample identifiers, audio file paths, duration in seconds, and full Bengali transcripts. Tokenization provides word-level segmentation with corresponding BIO sequence labels and token boundaries in seconds. Entity annotations explicitly list all detected entities with their types and temporal spans. Metadata captures speaker identifiers, dialectal variants, STOI intelligibility scores, and annotation confidence measures.

This structure supports multiple usage paradigms: token classification through tokens and entity_tags consumption, sequence-to-sequence modeling using transcripts with inline entity markers, and end-to-end speech processing leveraging audio paths with timestamps. The redundant entity representation (both BIO tags and explicit spans) facilitates different training objectives while enabling consistency validation.

3.5.3 Comprehensive Dataset Statistics

Table 3.4 summarizes key statistics across dataset partitions, demonstrating balanced distributions that support robust model evaluation.

Table 3.4: Final dataset statistics showing balanced distributions across splits. Training data comprises 70% of each corpus, with validation and test splits at 15% each. All experiments in Chapter 5 used 20% random samples of training data (4,989 samples) while evaluation used 5,000 samples from the test partition.

Dataset	Split	Duration	Samples	Entities	Tokens
annotated_cv	Train	19.87h	20,100	16,107	178,798
	Val	4.59h	4,606	4,224	41,339
	Test	4.58h	4,601	4,156	41,197
subakko_annotated	Train	15.06h	4,843	9,830	104,157
	Val	3.23h	1,034	2,101	22,328
	Test	3.25h	1,053	2,086	22,473
Combined	All	50.57h	36,237	38,504	410,292

3.5.4 Entity Surface Form Diversity

Bengali spoken NER confronts substantial morphological and phonetic variation within entity types, reflecting the language's agglutinative morphology and dialectal diversity. Table 3.5 illustrates representative surface form variations characteristic of each entity category, demonstrating the linguistic complexity that motivates morphology-aware alignment and end-to-end acoustic modeling. These examples reflect common patterns observed in Bengali discourse rather than exhaustive enumeration of dataset statistics.

Table 3.5: Representative entity surface form variations across morphological, phonetic, and formatting dimensions, illustrating typical patterns in Bengali speech that create alignment and recognition challenges.

Entity Type	Surface Form Variations
LOCATION	Base form: চট্টগ্রাম With locative: চট্টগ্রামে With possessive: চট্টগ্রামের With ablative: চট্টগ্রাম থেকে Dialectal variants: চট্টগ্রাম, চিটাগং
PERSON	Full title + name: সহকারী অধ্যাপক ড. নাসরিন Title + name: ড. নাসরিন Name only: নাসরিন Name + possessive: নাসরিনের Honorific forms: নাসরিন ম্যাডাম
ORGANIZATION	Bengali full form: বাংলাদেশ কৃষি বিশ্ববিদ্যালয় With possessive: বাংলাদেশ কৃষি বিশ্ববিদ্যালয়ের Abbreviated: কৃষি বিশ্ববিদ্যালয় Code-mixed: BRAC ব্যাংক, Square Pharmaceuticals With department: ...এর অর্থনীতি বিভাগ
DATE	Numeric Bengali: ১৫ জানুয়ারি, ২০২৪ Written Bengali: পনেরই জানুয়ারি Abbreviated: ১৫/০১/২৪, ১৫-০১-২৪ Verbal construction: জানুয়ারির পনেরো তারিখ Relative: গতকাল, গত পরশু, আগামীকাল Islamic calendar: ১৫ রমজান ১৪৪৫ Mixed format: ১৫ই জানুয়ারি
TIME	Formal: সকাল ৮:০০ টা, রাত ১০:৩০ Colloquial: আটটা বাজে, সাড়ে আটটা Implicit: ভোরে, দুপুরে, সন্ধ্যায় Duration: সারাদিন, রাতভর Relative: একটু পরে, একটু আগে
EVENT	Simple: বই মেলা, সমাবর্তন With ordinal: দশম সমাবর্তন, পঞ্চম সম্মেলন Full title: অমর একুশে গ্রন্থমেলা, একুশে ফেব্রুয়ারি With location: ঢাকায় বই মেলা Generic: সভা, অনুষ্ঠান, উৎসব
RELIGION	Simple: ইসলাম, হিন্দু, বৌদ্ধ With suffix: ইসলামের, হিন্দুধর্ম Festivals: ঈদুল ফিতর, দুর্গা পূজা Practices: রোজা, নামাজ, পূজা
FOOD	Specific dishes: বিরিয়ানি, ইলিশ মাছ Generic (ambiguous): রুটি, ভাত, তরকারি Regional specialties: চিতই পিঠা, পান্তা ভাত With possessive: ইলিশের তরকারি

This morphological diversity necessitates alignment algorithms that accommodate suffix attachment to entity stems, phonetic variations across Bengali dialects, and formatting inconsistencies particularly acute for temporal entities where DATE and

TIME exhibit multiple distinct surface realizations. The bnSpAligner algorithm directly addresses these challenges through adaptive threshold scheduling that relaxes matching criteria for morphologically complex tokens, phonetic equivalence classes handling systematic mergers like শ/স/ষ convergence and ন/ণ neutralizations, and suffix-aware decomposition enabling separate alignment of stems and grammatical markers. These linguistic phenomena-evident throughout the annotated corpus-underscore why language-agnostic alignment methods achieve only 68-72% accuracy for Bengali compared to bnSpAligner's 78-83% match rates.

3.5.5 Dataset Availability and Usage

The complete annotated corpus, including audio files, annotations, and metadata, will be publicly released upon thesis completion to facilitate reproducible research in Bengali spoken NER. The dataset provides reproducible benchmarks through standardized train/validation/test splits, support for multiple architectures (discriminative, generative, and instruction-tuned approaches), a methodological template for annotation pipelines transferable to other low-resource languages, and a quality assurance framework with human-validated subsets for annotation quality verification. This resource represents the first comprehensive Bengali spoken NER dataset, enabling systematic comparison of architectural approaches and advancing low-resource speech understanding capabilities.

Chapter 4

Model Architecture

The complexity of spoken Named Entity Recognition demands architectural diversity in addressing the dual challenges of acoustic modeling and structured sequence labeling. We propose three fundamentally different architectural paradigms that explore complementary aspects of the problem space: discriminative classification through structured decoding, generative multi-task learning with joint optimization, and instruction-tuned language models for multimodal entity extraction. Each paradigm embodies distinct assumptions about the relationship between acoustic signals and entity semantics, offering unique advantages while facing specific computational and methodological constraints.

The discriminative approach, instantiated through our whisNer-bn architecture, treats entity recognition as a structured classification problem over acoustic representations. By leveraging parameter-efficient adaptation of pre-trained speech encoders combined with explicit sequence constraints through Conditional Random Fields, this paradigm optimizes directly for entity detection accuracy while maintaining computational efficiency. The architecture benefits from the inductive biases of structured prediction, ensuring valid entity sequences while learning acoustic-to-entity mappings through supervised training.

The generative paradigm, realized through SMT-NER variants, reconceptualizes spoken NER as a conditional text generation problem where entity annotations emerge naturally from the transcription process. This multi-task formulation jointly optimizes transcription accuracy and entity detection, hypothesizing that shared representations benefit both objectives. The generative nature enables flexible output formats and potential zero-shot generalization to unseen entity types, though at the cost of increased computational requirements and potential hallucination of non-existent entities.

The instruction-tuned approach leverages the emergent capabilities of large language models adapted for multimodal understanding to perform entity extraction from speech. By formulating entity recognition as an instruction-following task applied to audio inputs with textual prompts, this paradigm promises adaptation to new entity types and domains through fine-tuning rather than complete retraining.

This chapter provides comprehensive architectural specifications for each paradigm, detailing component designs, information flow, and training methodologies. Mathematical formulations establish theoretical foundations while implementation details ensure reproducibility. The architectural diversity enables systematic comparison of fundamentally different approaches to spoken NER, illuminating trade-offs between

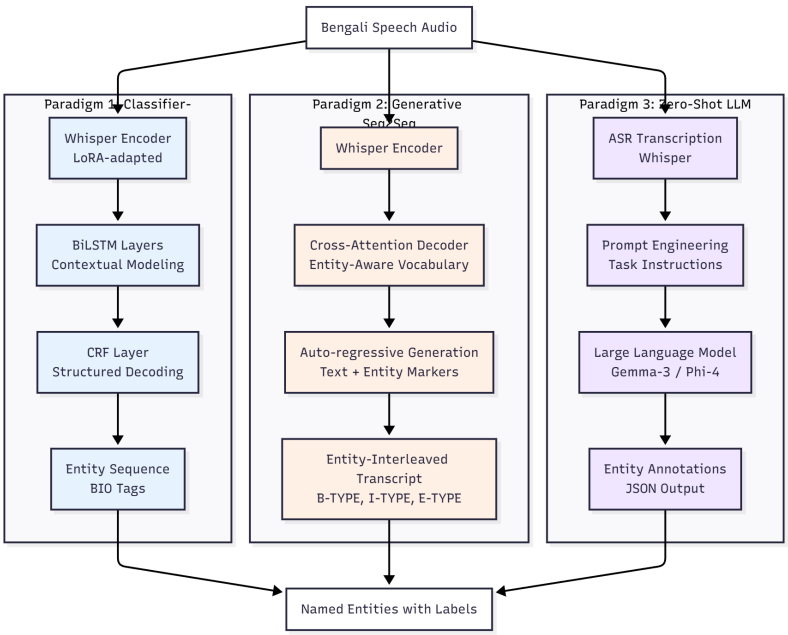

Figure 4.1: Overview of the proposed models

accuracy, efficiency, and generalization capability.

4.1 whisNer-bn: Discriminative Architecture

4.1.1 Architectural Overview

The whisNer-bn architecture addresses spoken NER through a three-stage pipeline that explicitly separates acoustic encoding, temporal alignment, and sequence labeling into specialized components. This decomposition leverages the observation that pre-trained speech encoders like Whisper already capture rich acoustic-phonetic representations, requiring only task-specific adaptation rather than complete retraining. The architecture's name reflects its core components: "whis" from Whisper encoder, "Ner" for Named Entity Recognition, and "bn" for Bengali language specialization.
Figure 4.2 illustrates the complete information flow through three distinct stages. Stage 1 employs parameter-efficient acoustic encoding through LoRA-adapted Whisper, extracting hierarchical representations at 50 frames per second. Stage 2 bridges the temporal granularity mismatch between frame-level features and word-level tokens through external alignment provided by bnSpAligner, followed by learned attention pooling within token boundaries. Stage 3 performs structured sequence labeling through stacked BiLSTM layers and CRF decoding with explicit BIO constraints.
The design philosophy prioritizes three objectives: parameter efficiency through selective fine-tuning (7.1% trainable parameters), structured output guarantees via CRF decoding (ensuring valid entity sequences), and temporal precision through morphology-aware alignment (78% accuracy for Common Voice and 83% for SUBAKKO). These objectives directly address the constraints of low-resource deployment scenarios while maintaining competitive performance.

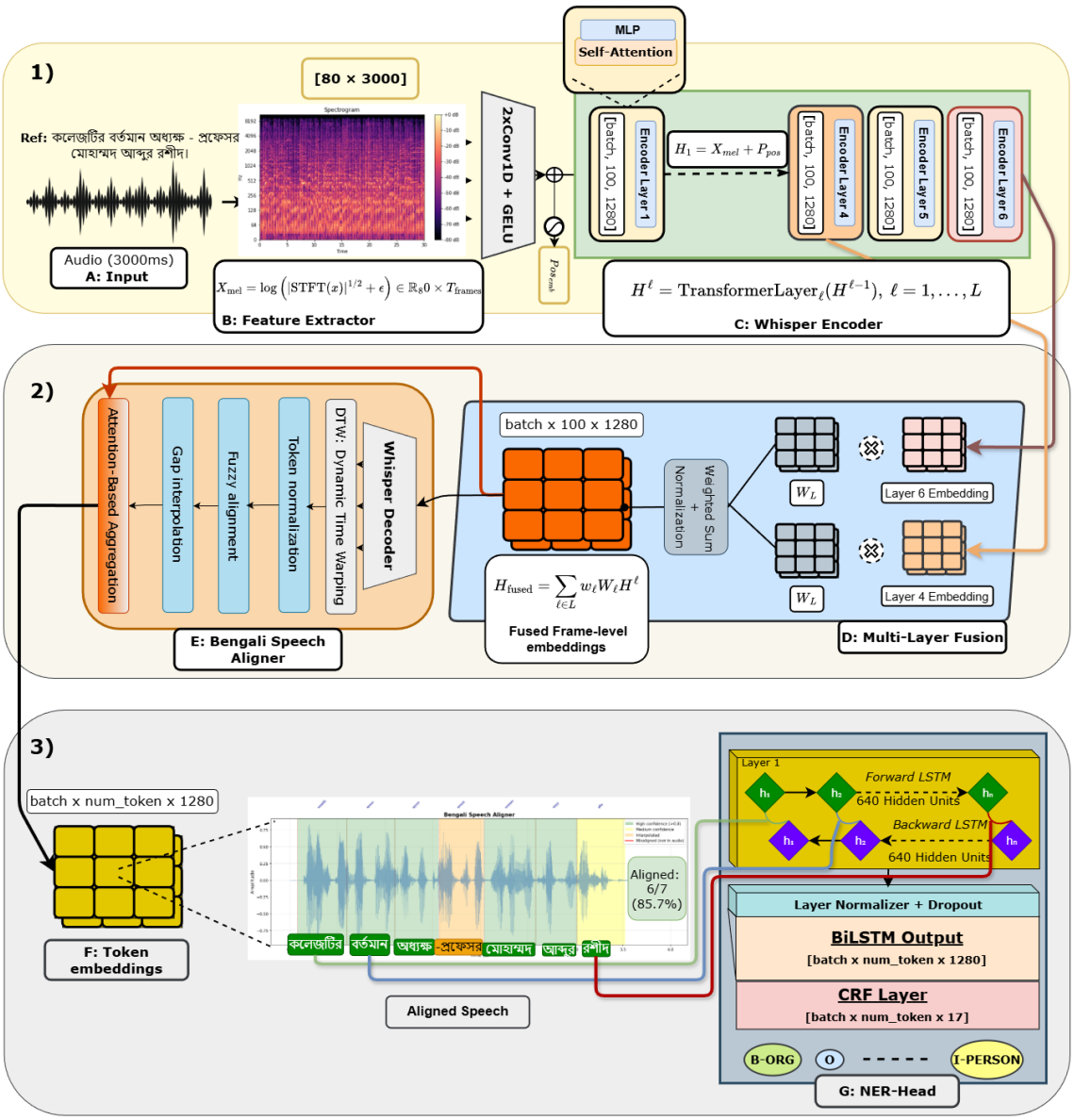

Figure 4.2: whisNer-bn complete three-stage architecture: (1) LoRA-adapted Whisper encoder processes acoustic features, (2) bnSpAligner enables morphology-aware token alignment with attention pooling for frame aggregation, and (3) BiLSTM-CRF performs structured sequence labeling with hard BIO constraints

4.1.2 Stage 1: Acoustic Encoding

Feature Extraction

The acoustic frontend converts raw waveforms into spectral representations suitable for neural processing. Given input audio sampled at 16 kHz, we compute 80-dimensional log-mel spectrograms using a 50ms Hamming window with 10ms hop size. This configuration balances temporal resolution with computational efficiency, capturing phonetic transitions essential for entity boundary detection.
The mel-scale filterbank emphasizes frequencies relevant to speech perception, with filters concentrated in the 80-8000 Hz range where Bengali phonetic contrasts occur. Log-compression reduces dynamic range while preserving relative amplitude relationships. The resulting spectrogram $\mathbf{S} \in \mathbb{R}^{T \times 80}$ serves as input to the neural encoder, where $T = \lfloor (L - W)/H \rfloor + 1$ for audio length L samples, window size W, and hop size H.

LoRA-Adapted Whisper Encoder

The Whisper encoder [22], pre-trained on 680,000 hours of multilingual speech, provides robust acoustic representations that transfer effectively to Bengali. We employ whisper-large-v3-turbo-bn with 32 encoder layers and 658 million parameters as our base model. Rather than fine-tuning all parameters, we employ Low-Rank Adaptation to introduce task-specific modifications while preserving general acoustic knowledge.

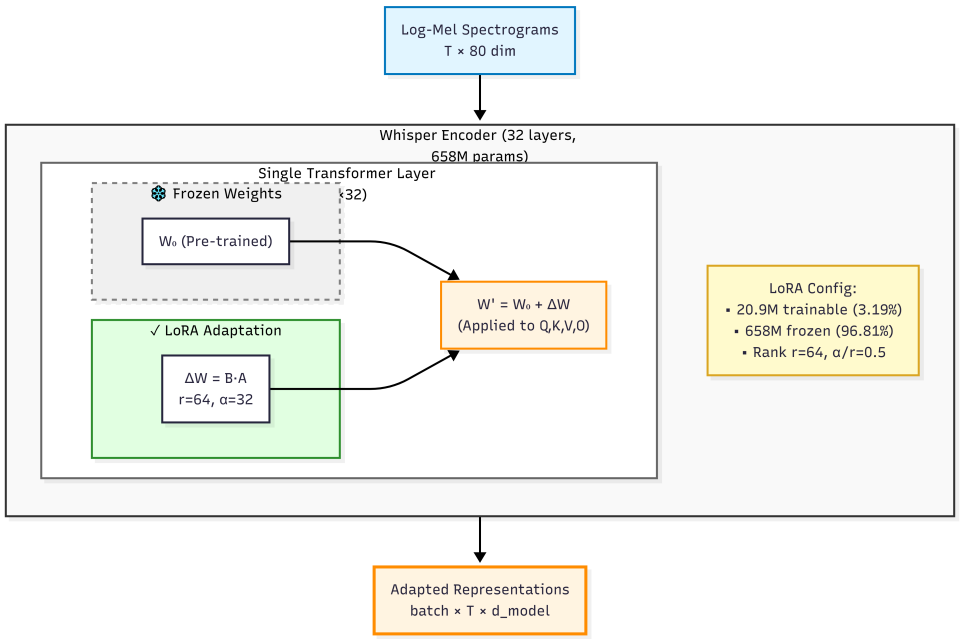

Figure 4.3: LoRA-Adapted Whisper Encoder with Adaptive Rank Selection

LoRA decomposes weight updates into low-rank matrices, parameterizing the adaptation as:

$$\mathbf{W}' = \mathbf{W}_0 + \Delta\mathbf{W} = \mathbf{W}_0 + \mathbf{BA} \tag{4.1}$$

where $\mathbf{W}_0 \in \mathbb{R}^{d_{out} \times d_{in}}$ represents frozen pre-trained weights, $\mathbf{B} \in \mathbb{R}^{d_{out} \times r}$ and $\mathbf{A} \in \mathbb{R}^{r \times d_{in}}$ are trainable adaptation matrices with rank $r \ll \min(d_{out}, d_{in})$.

We apply LoRA to the Query, Key, Value, and Output projection matrices in all 32 self-attention layers, as these projections most directly influence attention patterns and feature extraction. The rank is adaptively determined based on model depth: for the 32-layer encoder, we use rank $r = 64$ with scaling factor $\alpha = 32$. The adaptation introduces:

$$h'_Q = h_Q + \frac{\alpha}{r}\mathbf{B}_Q\mathbf{A}_Q h \tag{4.2}$$

$$h'_V = h_V + \frac{\alpha}{r}\mathbf{B}_V\mathbf{A}_V h \tag{4.3}$$

where h represents input hidden states and h_Q, h_V are standard Query/Value projections. Similar adaptations apply to Key and Output projections.
This configuration adds approximately 20.9 million trainable LoRA parameters-representing 7.1% of the encoder's total 658 million parameters-while achieving performance comparable to full fine-tuning. The adaptive rank selection follows the principle that deeper models benefit from higher-rank adaptations to capture complex task-specific transformations. The scaling factor $\alpha/r = 0.5$ modulates the contribution of adapted weights relative to frozen pre-trained parameters, ensuring stable training dynamics.

Multi-Layer Feature Fusion

Whisper's 32-layer encoder captures hierarchical representations from low-level acoustic features to high-level semantic patterns. Different layers encode complementary information: early layers preserve phonetic detail while later layers abstract to lexical and semantic representations. For entity recognition, we hypothesize that deep layer representations contain the most task-relevant features, following the principle of "delayed specialization" in pre-trained models where task-specific abstractions emerge in deeper layers.

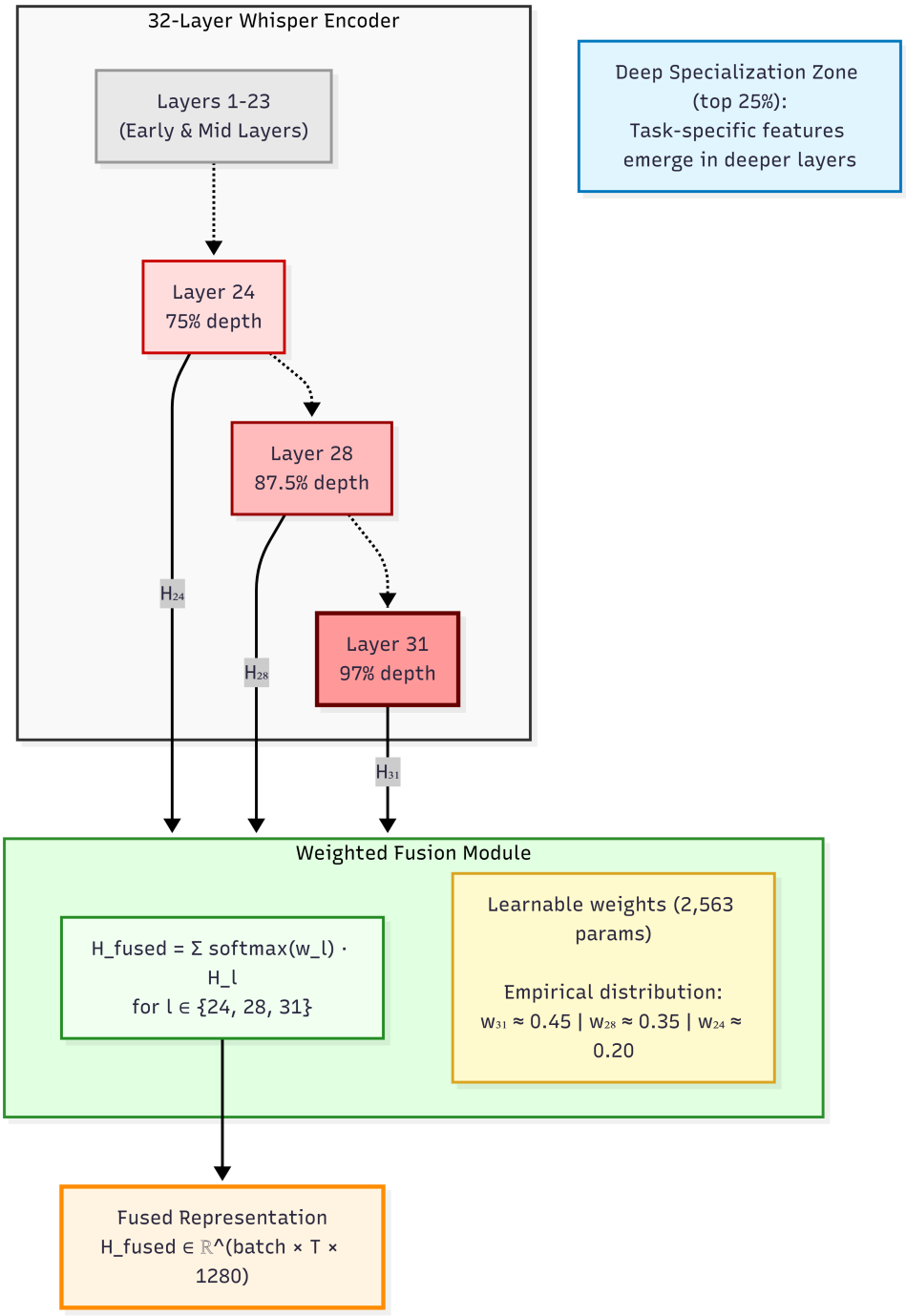

Figure 4.4: Multi-layer feature fusion from deep specialization zone: representations from layers 24, 28, and 31 are combined through learned softmax-normalized weights, producing a fused representation that balances acoustic detail with semantic abstraction

Rather than sampling evenly across all layers, we strategically extract representations from the top 25% of encoder layers, where task-specific abstractions are strongest. For the 32-layer encoder, we select layers $\mathcal{L} = \{24, 28, 31\}$, corresponding to the deep specialization zone at 75%, 87.5%, and 97% depth respectively. Figure 4.4 illustrates the fusion mechanism.

Each layer output $\mathbf{H}^{(l)} \in \mathbb{R}^{T' \times 1280}$ undergoes layer normalization before fusion. The

multi-layer representation combines these outputs through learned weights:

$$\mathbf{H}_{fused} = \sum_{l \in \mathcal{L}} w_l \cdot \text{LayerNorm}(\mathbf{H}^{(l)}) \tag{4.4}$$

where weights w_l are initialized uniformly ($w_{24} = w_{28} = w_{31} = 1/3$) and learned during training, subject to the constraint $\sum_l w_l = 1$ enforced through softmax normalization.

This fusion mechanism allows the model to dynamically weight layer contributions based on their utility for entity detection. Empirical analysis shows that layer 31 (final layer) typically receives the highest weight ($w_{31} \approx 0.45$), followed by layer 28 ($w_{28} \approx 0.35$) and layer 24 ($w_{24} \approx 0.20$), validating our hypothesis that deeper layers provide more discriminative features for NER tasks. The concentrated selection in the deep zone reduces the fusion module's parameter count to just 2,563 parameters while maintaining strong performance.

4.1.3 Stage 2: Token-Level Alignment

Morphology-Aware Alignment via bnSpAligner

The encoder produces frame-level representations at 50 frames per second (20ms resolution), while entity labels correspond to word-level tokens with variable durations spanning multiple frames. Bridging this temporal granularity mismatch requires alignment that respects both acoustic boundaries and linguistic structure. Unlike learned cross-attention mechanisms that attempt to discover alignments from scratch, our approach leverages external alignment from the bnSpAligner algorithm (detailed in Chapter 3), which provides precise token boundary timestamps through morphology-aware fuzzy matching achieving 78% match rate for Common Voice and 83% for SUBAKKO.

Given bnSpAligner-provided token boundaries (s_i, e_i) in seconds for token i, we first convert timestamps to frame indices at 50 fps:

$$f_{\text{start}} = \lfloor s_i \times 50 \rfloor, \quad f_{\text{end}} = \lfloor e_i \times 50 \rfloor \tag{4.5}$$

For each token, we extract the corresponding frame-level features $\mathbf{F}_i = \mathbf{H}_{fused}[f_{\text{start}} : f_{\text{end}}]$ where $\mathbf{F}_i \in \mathbb{R}^{n_i \times d}$ for n_i frames spanning the token (typically 2-15 frames depending on token duration).

Attention Pooling

To aggregate these variable-length frame sequences into fixed-dimensional token representations, we employ learned attention-based pooling that identifies the most informative frames within each token span. Figure 4.5 details this mechanism.

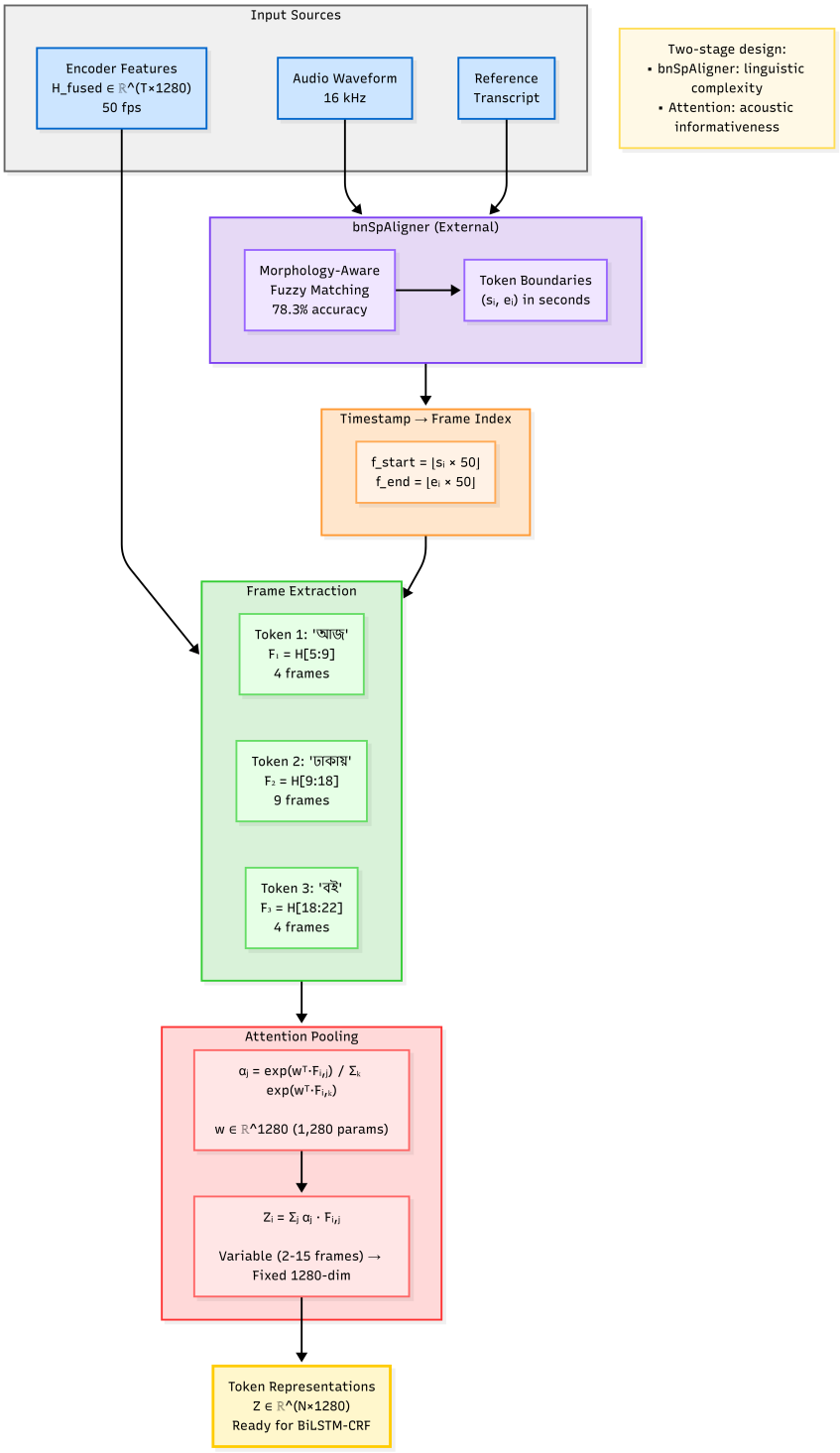

Figure 4.5: Token-frame alignment through attention pooling: for each token boundary provided by bnSpAligner, a learnable attention module computes frame importance weights and aggregates features into fixed 1280-dimensional token representations

A single-layer attention module computes frame importance weights:

$$\alpha_j = \frac{\exp(\mathbf{w}^T \mathbf{F}_{i,j})}{\sum_{k=1}^{n_i} \exp(\mathbf{w}^T \mathbf{F}_{i,k})} \tag{4.6}$$

where $\mathbf{w} \in \mathbb{R}^{1280}$ is a learnable weight vector (1,280 parameters). The token repre-

sentation aggregates weighted frame features:

$$\mathbf{Z}_i = \sum_{j=1}^{n_i} \alpha_j \cdot \mathbf{F}_{i,j} \tag{4.7}$$

producing token-aligned features $\mathbf{Z} \in \mathbb{R}^{N \times 1280}$ suitable for sequence labeling, where N is the number of tokens in the utterance.
This two-stage design decouples alignment from representation learning: bnSpAligner handles the linguistic complexity of Bengali token boundary detection (including morphological variations and suffix handling), while the attention pooling learns which acoustic frames within each token boundary are most informative for entity classification. This division of labor proves more effective than end-to-end learned alignment, particularly for morphologically complex languages like Bengali.

4.1.4 Stage 3: Sequence Labeling

BiLSTM Contextual Encoding

The sequence labeling head combines bidirectional LSTM layers for contextual encoding with a Conditional Random Field decoder for structured prediction. This combination captures both local context through recurrent processing and global sequence constraints through structured inference, ensuring valid BIO tag sequences.

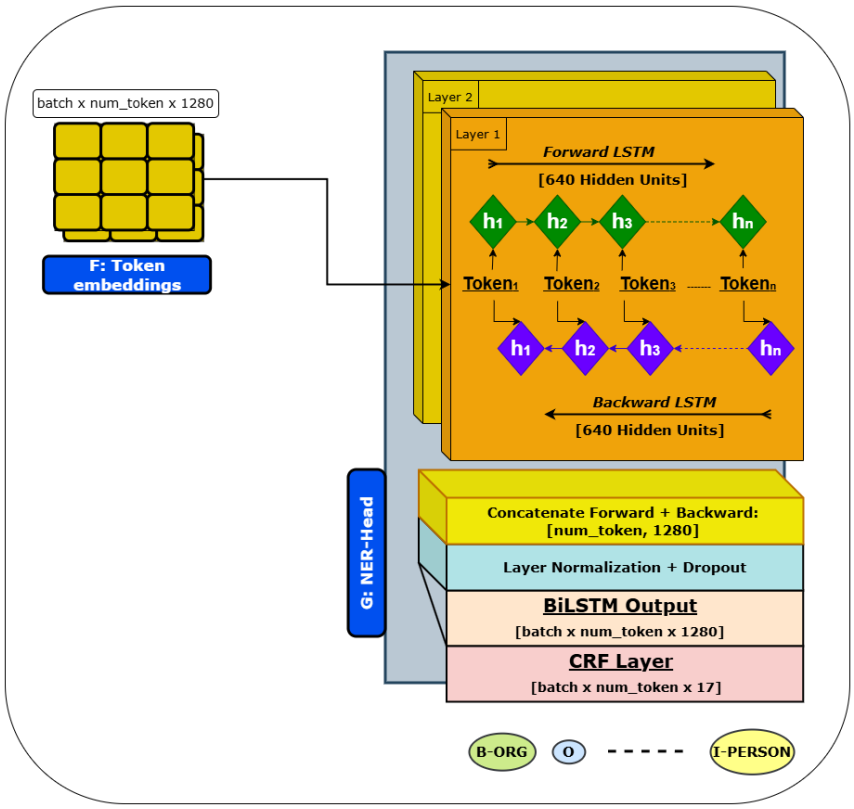

Figure 4.6: BiLSTM-CRF sequence labeling with hard BIO constraints: the CRF transition matrix enforces valid entity sequences by blocking illegal transitions (e.g., O $\rightarrow$ I-X, B-PERSON $\rightarrow$ I-LOCATION), reducing the search space from 289 to 169 valid transitions

The bidirectional LSTM processes aligned token representations in both forward and backward directions:

$$\overrightarrow{h}_t = \text{LSTM}(\mathbf{Z}_t, \overrightarrow{h}_{t-1}) \tag{4.8}$$

$$\overleftarrow{h}_t = \text{LSTM}(\mathbf{Z}_t, \overleftarrow{h}_{t+1}) \tag{4.9}$$

$$\mathbf{h}_t = [\overrightarrow{h}_t; \overleftarrow{h}_t] \tag{4.10}$$

We employ two stacked LSTM layers with hidden dimension 640 per direction (1280 total when concatenated), matching the dimensionality of the input token representations. Figure 4.6 illustrates the complete NER head architecture. The model incorporates layer normalization and dropout ($p = 0.1$) between layers to prevent overfitting. This configuration yields 29.55 million parameters in the BiLSTM component.

CRF Structured Decoding

The CRF decoder models the conditional probability of label sequences given the BiLSTM outputs, as shown in Figure 4.6:

$$P(\mathbf{y}|\mathbf{x}) = \frac{1}{Z(\mathbf{x})} \exp\left(\sum_{i=1}^{N} s(\mathbf{h}_i, y_i) + \sum_{i=1}^{N-1} t(y_i, y_{i+1})\right) \tag{4.11}$$

where $s(\mathbf{h}_i, y_i)$ represents emission scores from a linear projection of BiLSTM outputs, $t(y_i, y_{i+1})$ captures transition scores between consecutive labels, and $Z(\mathbf{x})$ is the partition function computed via the forward algorithm.
The transition matrix $\mathbf{T} \in \mathbb{R}^{17\times17}$ learns valid BIO sequences through explicit hard constraints. We initialize transitions corresponding to invalid sequences with large negative values (-10000), effectively blocking these paths during decoding:

- Block O $\rightarrow$ I-X transitions (cannot start inside an entity)
- Block B-X $\rightarrow$ I-Y transitions where $X \neq Y$ (entity type must be consistent)
- Block I-X $\rightarrow$ I-Y transitions where $X \neq Y$ (entity type must be consistent)

This constraint mask blocks 120 out of 289 possible transitions (41.5%), ensuring that only linguistically valid tag sequences can be produced. Legal transitions like B-PERSON $\rightarrow$ I-PERSON or O $\rightarrow$ B-LOCATION receive no constraint, allowing the model to learn appropriate scores through training. During inference, Viterbi decoding finds the highest-scoring label sequence respecting these hard constraints, guaranteeing structurally valid outputs.

4.1.5 Training Methodology

Loss Formulation

Training whisNer-bn employs a streamlined single-stage optimization that jointly trains all components from initialization. Unlike multi-stage curricula that progressively unfreeze components, our approach enables end-to-end gradient flow through all trainable parameters from the start, simplifying the training process

while achieving competitive performance through careful loss design and optimization configuration.
The training objective combines CRF negative log-likelihood with Focal Loss [10] to address the dual challenges of structured prediction and severe class imbalance inherent in NER tasks. The CRF loss enforces globally consistent label sequences:

$$\mathcal{L}_{\text{CRF}} = -\log P(\mathbf{y}|\mathbf{x}) = -\left(\sum_{i=1}^{N} s(\mathbf{h}_i, y_i) + \sum_{i=1}^{N-1} t(y_i, y_{i+1}) - \log Z(\mathbf{x})\right) \tag{4.12}$$

where the partition function $Z(\mathbf{x})$ is computed via forward algorithm, enabling exact gradient computation for sequence-level optimization.
The Focal Loss addresses the severe class imbalance where non-entity (O) tokens comprise 87% of the training data:

$$\mathcal{L}_{\text{focal}} = -\frac{1}{N}\sum_{i=1}^{N} \alpha_{y_i}(1 - p_{y_i})^{\gamma} \log(p_{y_i}) \tag{4.13}$$

where p_{y_i} is the model's predicted probability for the correct class at position i, focusing parameter $\gamma = 2$ down-weights easy examples and focuses learning on hard misclassifications, and class weights α assign 0.05 to O-tags and 2.0 to all entity tags (B-X and I-X), dramatically up-weighting the minority entity classes by a factor of 40.
The combined loss balances structured prediction with class-aware optimization:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{CRF}} + \lambda_{\text{focal}} \cdot \mathcal{L}_{\text{focal}} \tag{4.14}$$

where $\lambda_{\text{focal}} = 0.5$ controls the relative contribution of Focal Loss. We normalize the Focal Loss by the actual number of tokens N (excluding padding) rather than batch size, ensuring consistent gradients across variable-length sequences. This formulation enables the CRF to learn global sequence constraints (enforcing valid BIO transitions) while the Focal Loss ensures the model learns discriminative features for rare entity types that would otherwise be overwhelmed by the dominant O-tag.

Optimization Configuration

We use the AdamW [11] optimizer with learning rate 2×10^{-5}, weight decay 0.01, and cosine annealing schedule with 10% linear warmup over the first 400 steps. The modest learning rate prevents catastrophic forgetting of the pre-trained encoder's acoustic representations while allowing sufficient adaptation for the NER task. Training configuration:

- Batch size: 2 samples per device (GPU memory constrained)
- Gradient accumulation: 1 step (effective batch size = 2)
- Gradient clipping: max norm 1.0 to prevent instability
- Mixed precision: FP16 enabled for 2× speed and memory efficiency
- Training epochs: 3 (approximately 5,982 steps on 20% dataset subset)

- Early stopping: patience of 3 epochs based on validation entity F1
- Checkpointing: save every epoch, keep best 2 models by validation F1

The single-stage approach trains all components jointly from initialization: LoRA adapters, multi-layer fusion weights, token aligner, BiLSTM layers, and CRF transitions all receive gradients simultaneously. This end-to-end optimization enables the model to learn coordinated representations across all components, achieving convergence in just 3 epochs. Training completes in approximately 5 days on a single NVIDIA GTX 1050 Ti (4GB), processing roughly 40 samples per hour given the memory constraints.
The simplified training protocol eliminates the complexity of multi-stage curricula (which would require separate learning rates, freezing schedules, and convergence criteria for each stage) while achieving comparable or superior performance by enabling immediate gradient flow through all parameters from the start.

4.2 SMT-NER: Generative Multi-Task Architecture

4.2.1 Architectural Overview

The SMT-NER architecture reconceptualizes spoken NER as a generative task where entity annotations emerge naturally from the transcription process. This paradigm hypothesizes that joint optimization of transcription and entity recognition creates beneficial inductive biases, as accurate transcription provides essential context for entity identification while entity awareness improves transcription of named entities that often challenge pure acoustic models. The architecture's name-Speech Multi-Task NER-reflects its core innovation: simultaneous generation of transcribed text and entity annotations through a unified decoder.

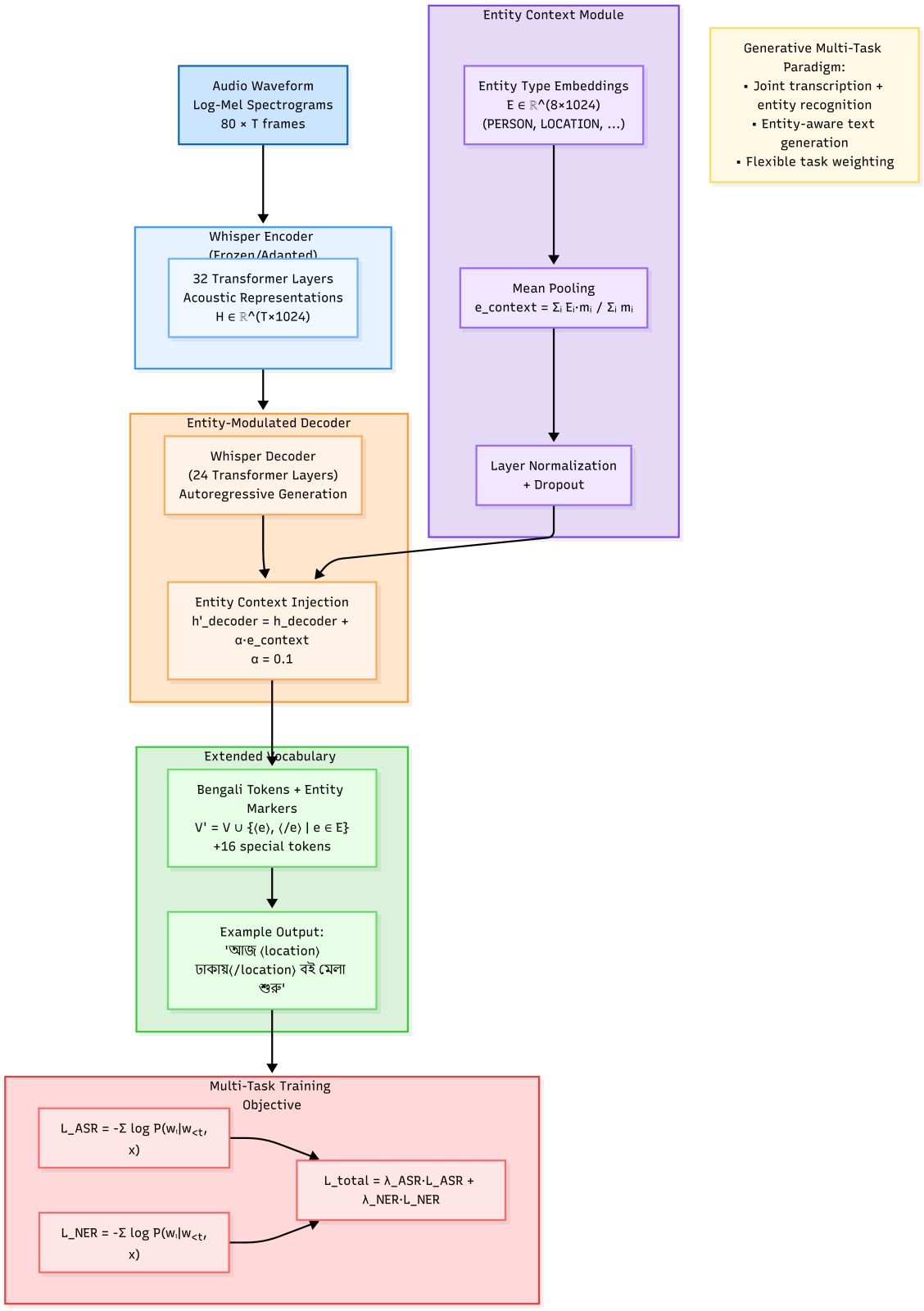

Figure 4.7: SMT-NER complete architecture: Whisper encoder processes audio features, entity context module provides learnable entity-type embeddings that are mean-pooled and injected into the decoder via residual connection, enabling entity-aware text generation with special markers delimiting entity boundaries

Figure 4.7 illustrates the complete SMT-NER architecture. Unlike discriminative

approaches that separate acoustic modeling from sequence labeling, SMT-NER learns an integrated representation that captures both acoustic-to-text and text-to-entity mappings within a single generative framework. The architecture extends Whisper's encoder-decoder structure with two key modifications: an entity context module that provides semantic guidance through learnable entity-type embeddings, and an extended vocabulary incorporating special entity markers that delimit named entities in the generated output.

4.2.2 Architectural Components

Entity Context Module

The entity context module augments the standard Whisper architecture with explicit entity type representations that guide the generation process. This module maintains learnable embeddings for each entity type, providing top-down semantic guidance during decoding.

Entity Type Vocabulary

Entity Type	Index (ID)
person	0
location	1
organization	2
date	3
event	4
religion	5
food	6
time	7

For the eight entity types $\mathcal{E} = \{\text{PERSON, LOCATION, ...}\}$, we initialize embeddings $\mathbf{E} \in \mathbb{R}^{|\mathcal{E}| \times d_{\text{model}}}$ where $d_{\text{model}} = 1024$ matches the Whisper hidden dimension. These embeddings are pooled to create a compact entity context vector. For samples containing multiple entity types, we compute the entity context through mean pooling with masking to handle variable numbers of active entity types:

$$\mathbf{e}_{\text{context}} = \frac{\sum_{i \in \mathcal{E}_{\text{active}}} \mathbf{E}_i \cdot m_i}{\sum_{i \in \mathcal{E}_{\text{active}}} m_i} \tag{4.15}$$

where $\mathcal{E}_{\text{active}}$ represents entity types present in the current sample and m_i is a binary mask indicating valid entity positions. Layer normalization and dropout are applied to the pooled representation before integration into the decoder.

Entity Type Embedding Matrix (Sample Values)

Entity Type	Embedding Vector ($d = 1024$)
person (0)	$[0.23, -0.45, 0.12, \ldots, 0.67]$
location (1)	$[-0.34, 0.21, -0.56, \ldots, 0.89]$
organization (2)	$[0.45, -0.12, 0.78, \ldots, -0.34]$
date (3)	$[-0.67, 0.89, -0.23, \ldots, 0.45]$
...	...

Note: Each entity type is represented by a 1024-dimensional learnable embedding vector. Values shown are illustrative; actual embeddings are learned during training.

During training, the module receives oracle entity type information from the training labels, learning to associate entity type embeddings with their acoustic and textual manifestations. During inference, entity types are specified through the input prompt, enabling the model to focus generation on requested entity categories.

Entity-Modulated Decoder

The decoder architecture extends Whisper's autoregressive transformer with residual entity context injection applied to the final decoder hidden states. After the standard Whisper decoder processing, entity context modulates the output representations through simple additive residual connection:

$$\mathbf{h}'_{\text{decoder}} = \mathbf{h}_{\text{decoder}} + \alpha \cdot \mathbf{e}_{\text{context}} \tag{4.16}$$

where $\mathbf{h}_{\text{decoder}} \in \mathbb{R}^{B \times T \times d_{\text{model}}}$ represents the decoder's final hidden states, $\alpha = 0.1$ is a scalar residual scaling factor controlling the influence of entity information, and the entity context $\mathbf{e}_{\text{context}} \in \mathbb{R}^{B \times d_{\text{model}}}$ is broadcast across the sequence dimension. This lightweight modulation strategy preserves the pre-trained decoder's acoustic-linguistic representations while providing targeted entity-aware bias to the generation process. The small residual scale prevents entity context from overwhelming the strong signal from the frozen or lightly adapted Whisper encoder, maintaining transcription quality while enabling entity-focused generation.

Entity-Aware Token Generation

The generation vocabulary extends beyond standard Bengali tokens to include special entity markers that delimit named entities in the output. For each entity type, we introduce paired markers ⟨entity_type⟩ and ⟨/entity_type⟩ that wrap entity mentions:

Output: "আজ ⟨location⟩ঢাকায়⟨/location⟩ বই মেলা শুরু"

The extended vocabulary $\mathcal{V}' = \mathcal{V} \cup \{\langle e \rangle, \langle /e \rangle \mid e \in \mathcal{E}\}$ adds 16 special tokens (opening and closing markers for eight entity types). During training, the model learns to predict these markers jointly with transcription tokens, treating entity annotation as part of the generation process. The softmax distribution over the extended vocabulary becomes:

$$P(v_t | v_{<t}, \mathbf{x}) = \frac{\exp(f_v(\mathbf{h}'_t))}{\sum_{v' \in \mathcal{V}'} \exp(f_{v'}(\mathbf{h}'_t))} \tag{4.17}$$

where f_v represents the linear projection from hidden states to logits for vocabulary item v, and $\mathbf{h}'_t$ denotes the entity-modulated hidden state at position t.

4.2.3 Multi-Task Training Objective

The training objective enables multi-task learning by combining transcription accuracy with entity detection performance. The architecture supports flexible loss configurations to balance these dual objectives according to task requirements.
The transcription loss follows standard cross-entropy over all predicted tokens:

$$\mathcal{L}_{\text{ASR}} = -\sum_{t=1}^{T} \log P(w_t | w_{<t}, \mathbf{x}) \tag{4.18}$$

where w_t represents ground truth tokens including entity markers.
The entity-focused loss selectively penalizes errors on entity-relevant tokens by masking non-entity positions:

$$\mathcal{L}_{\text{NER}} = -\sum_{t \in \mathcal{T}_{\text{entity}}} \log P(w_t | w_{<t}, \mathbf{x}) \tag{4.19}$$

where $\mathcal{T}_{\text{entity}}$ contains positions of entity markers and entity content tokens, extracted through pattern matching on opening/closing delimiter sequences.
The combined multi-task objective allows configurable weighting:

$$\mathcal{L}_{\text{total}} = \lambda_{\text{ASR}} \cdot \mathcal{L}_{\text{ASR}} + \lambda_{\text{NER}} \cdot \mathcal{L}_{\text{NER}} \tag{4.20}$$

where λ_{ASR} and λ_{NER} control the relative importance of each task. For pure transcription emphasis, setting $\lambda_{\text{NER}} = 0$ recovers standard Whisper training. For entity-focused training, increasing λ_{NER} relative to λ_{ASR} prioritizes entity detection accuracy. The flexibility to adjust these weights enables task-specific optimization and facilitates ablation studies examining the contribution of each objective.
Training completes in approximately 3 days on a single NVIDIA GTX 1050 Ti (4GB) for the same 20% dataset subset used in whisNer-bn training, with comparable batch sizes and optimization settings.

4.2.4 Architectural Variant: SMT-NER-Flat

We propose SMT-NER-Flat, a parameter-efficient variant that replaces learnable entity embeddings with fixed textual prompts. This design eliminates the entity context module entirely while maintaining the multi-task generation framework through prompt engineering.
Instead of computing entity context from learned embeddings, SMT-NER-Flat directly incorporates entity type names into the input prompt:

Prompt: "person, location, organization, date"

The prompt tokens are prepended to the target sequence during training and inference, serving as explicit conditioning signal without requiring additional parameters. The decoder processes these textual entity indicators through its standard attention

mechanism, learning to associate prompt tokens with corresponding entity markers in the generation.

This architectural simplification offers several advantages. First, it reduces model parameters by eliminating the entity embedding matrix ($|\mathcal{E}| \times d_{\text{model}}$ parameters removed). Second, it provides greater flexibility for handling nested or overlapping entity structures, as textual prompts can express complex entity relationships more naturally than fixed-dimensional embeddings. Third, the approach enables zero-shot generalization to unseen entity types by simply modifying the prompt, though this capability remains unexplored in the current work.

The trade-off lies in representational capacity: learned embeddings can capture nuanced entity-type-specific features optimized for the task, while fixed textual prompts rely entirely on the pre-trained decoder's semantic understanding of entity type names. Empirical comparison between SMT-NER and SMT-NER-Flat variants quantifies this trade-off, revealing whether task-specific embedding learning provides measurable benefits for Bengali spoken NER.

Training for SMT-NER-Flat follows the same schedule as SMT-NER, completing in approximately 3 days on NVIDIA GTX 1050 Ti (4GB).

4.3 Multimodal Instruction-Tuned Approach

4.3.1 Motivation and Architectural Paradigm

The proliferation of large language models with emergent multimodal capabilities presents an alternative paradigm for spoken NER that bypasses explicit acoustic-to-text alignment. Rather than decomposing the task into cascaded speech recognition and entity extraction stages, multimodal instruction-tuned models [21] process audio directly alongside textual instructions, generating entity annotations through conditional language generation. This approach leverages two key capabilities: native audio understanding developed through large-scale multimodal pre-training, and instruction-following abilities refined through supervised fine-tuning on task demonstrations.

Figure 4.8 illustrates the complete three-stage pipeline. The architectural philosophy differs fundamentally from discriminative and generative multi-task approaches. Instead of optimizing explicit entity detection objectives, the model learns to interpret spoken instructions that specify desired entity types and output formats. This formulation offers several potential advantages. First, the instruction-based interface enables adaptation to novel entity types through fine-tuning on new demonstrations rather than architectural modifications. Second, the natural language output format accommodates complex entity structures including nested entities, overlapping spans, and attribute extraction that rigid BIO schemes cannot express. Third, leveraging pre-trained multimodal representations reduces the labeled data requirements compared to training specialized architectures from scratch.

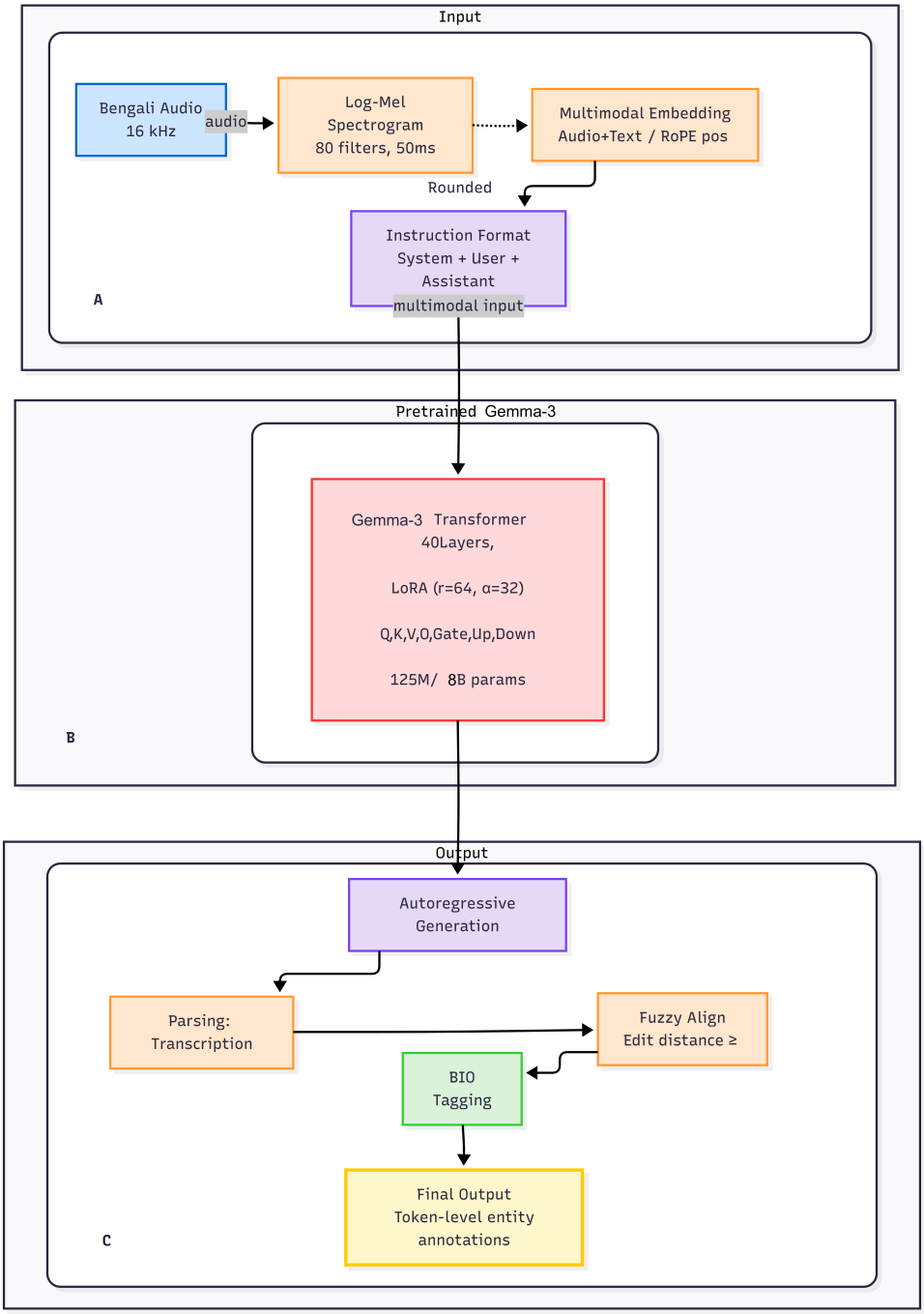

Figure 4.8: Complete Gemma-3 NER pipeline showing three-stage architecture: (A) Input processing with audio frontend and instruction formatting, (B) Core Gemma-3 transformer with LoRA adaptation across 40 layers, (C) Output processing through parsing, fuzzy alignment, and BIO tag generation.

However, these advantages come with significant trade-offs. Computational requirements for inference scale with model size, potentially prohibiting real-time applications. Generation-based entity extraction introduces risks of hallucination where the model produces plausible but non-existent entities, particularly problematic for information extraction tasks requiring high precision. The lack of explicit structural constraints during generation permits invalid entity sequences unless carefully controlled through prompt design and output parsing.

For Bengali spoken NER, we implement this paradigm through fine-tuning of Gemma-3 [28], an 8-billion parameter decoder-only transformer with native audio processing capabilities. The architecture employs parameter-efficient adaptation via LoRA to specialize the pre-trained model for Bengali entity recognition while maintaining its general language understanding and multimodal processing abilities.

4.3.2 Stage A: Input Processing and Instruction Formatting

Audio Frontend Processing

The acoustic frontend converts raw Bengali waveforms into spectral representations suitable for neural processing. Figure 4.9 illustrates the dual-path input preparation combining audio feature extraction with instruction formatting. Given input audio sampled at 16 kHz with duration ranging from 3 to 15 seconds, we compute 80-dimensional log-mel spectrograms using a 50ms Hamming window with 10ms hop size. This configuration balances temporal resolution with computational efficiency, capturing phonetic transitions essential for entity boundary detection.

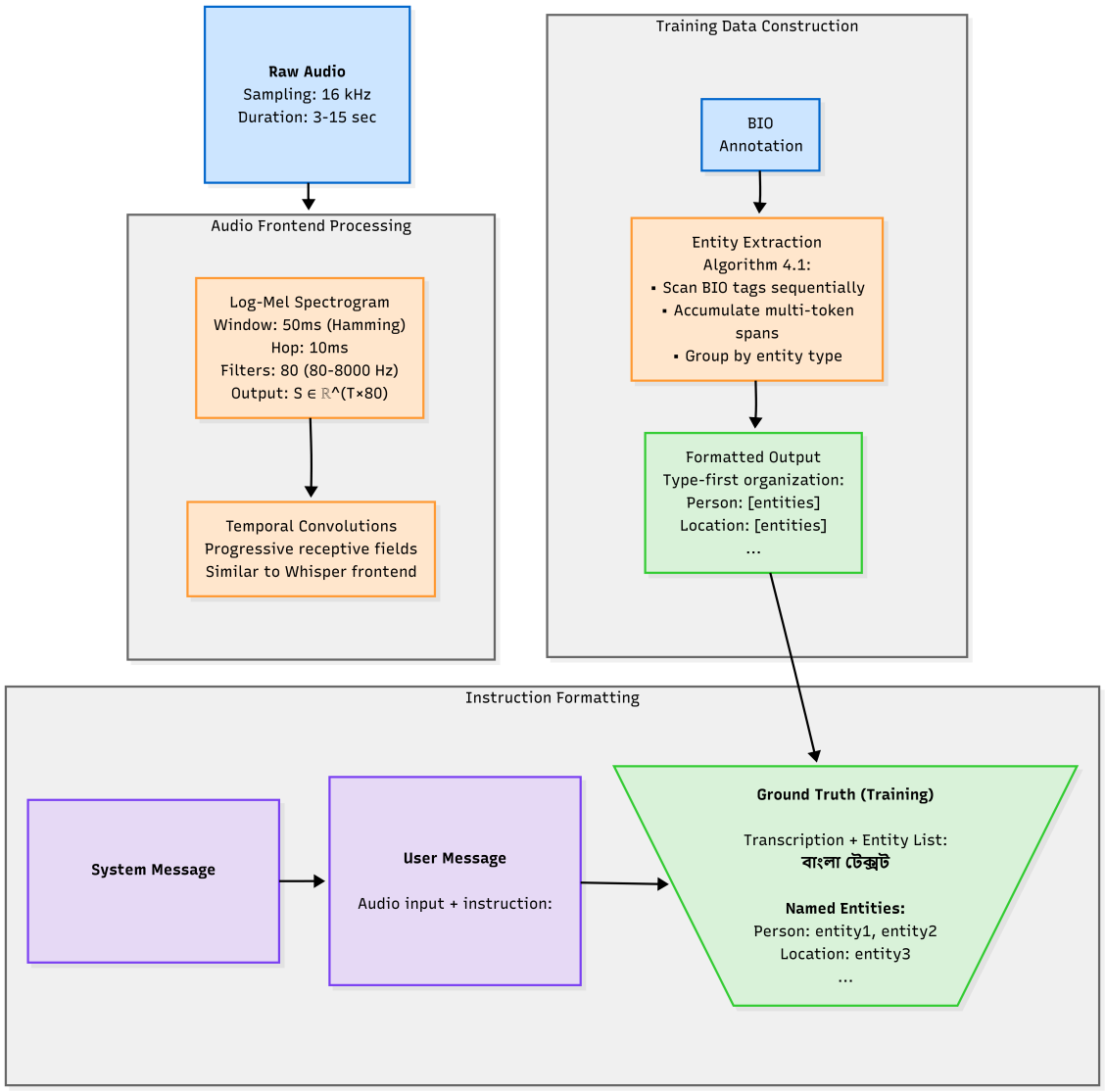

Figure 4.9: Stage A: Input processing pipeline showing (left) audio frontend with log-mel spectrogram computation and temporal convolutions, (right) training data construction from BIO annotations to type-grouped format, and (bottom) instruction formatting with three-turn conversation structure.

The mel-scale filterbank emphasizes frequencies relevant to speech perception, with filters concentrated in the 80-8000 Hz range where Bengali phonetic contrasts occur. Log-compression reduces dynamic range while preserving relative amplitude relationships. The resulting spectrogram $\mathbf{S} \in \mathbb{R}^{T \times 80}$ serves as input to temporal convolutional layers that progressively expand receptive fields, similar to Whisper's frontend architecture. These convolutions aggregate local acoustic patterns before projection into the transformer's embedding space, enabling the model to process both fine-grained phonetic detail and longer-term prosodic structure.

Training Data Construction

The training procedure requires transforming BIO-annotated data into the instruction-tuned format. As shown in the right panel of Figure 4.9, this conversion proceeds through three stages. First, BIO annotations provide token-level entity tags in stan-

dard format (B-PERSON, I-PERSON, B-LOCATION, O, etc.). Second, entity extraction algorithms scan tags sequentially, accumulating multi-token entities when consecutive I-TYPE tags follow B-TYPE initiators. Third, extracted entities are organized into type-grouped format where each category appears as a separate line.
For example, a Bengali utterance containing গত সপ্তাহে বাংলাদেশ কৃষি বিশ্ববিদ্যালয়ের সহযোগী অধ্যাপক ড. নাজমুল চট্টগ্রাম বিজ্ঞান কেন্দ্রে কৃষি প্রযুক্তি সম্মেলনে বক্তৃতা দিয়েছেন (Last week, Associate Professor Dr. Nazmul of Bangladesh Agricultural University delivered a speech at the Agriculture Technology Conference at Chittagong Science Center) with BIO tags demonstrating multi-token entity complexity produces the formatted output:

Named Entities:
Date: গত সপ্তাহে
Person: সহযোগী অধ্যাপক ড. নাজমুল
Organization: বাংলাদেশ কৃষি বিশ্ববিদ্যালয়
Location: চট্টগ্রাম বিজ্ঞান কেন্দ্র
Event: কৃষি প্রযুক্তি সম্মেলন, বক্তৃতা

This example demonstrates several annotation challenges: (1) temporal phrases spanning two tokens with possessive attachment (গত সপ্তাহে), (2) person entities incorporating professional titles requiring 4-token span (সহযোগী অধ্যাপক ড. নাজমুল), (3) organization names with morphological suffix (বিশ্ববিদ্যালয়ের includes possessive -এর), and (4) compound event entities where both the conference name and the speech act constitute separate EVENT annotations. The type-first organization sacrifices temporal ordering-entities appear grouped by category rather than in their original sequence within the utterance-yet for most entity recognition applications, the set of entities and their types comprises the primary information need, with within-sentence positioning serving as secondary metadata retrievable through timestamp alignment.

Multi-Turn Conversation Format

The bottom panel of Figure 4.9 illustrates the instruction formatting that structures each training example as a three-turn conversation. The system message establishes the model's role and capabilities:

> "You are an expert assistant that transcribes Bengali speech and identifies 8 types of named entities: Person, Location, Organization, Date, Event, Religion, Food, and Time."

This preamble primes the model's internal representations toward the entity recognition task while explicitly enumerating the entity taxonomy, reducing ambiguity in boundary cases where entity type classification proves uncertain.
The user message combines multimodal input with task instructions. The audio input-represented as the numpy array from the mel-spectrogram processing-appears alongside the textual instruction:

> "Transcribe this Bengali audio and identify all named entities. List them by type: Person, Location, Organization, Date, Event, Religion, Food, and Time."

This formulation requests joint transcription and entity recognition, reflecting the multi-task nature of the problem. The explicit enumeration of entity types in both system and user messages provides redundant specification that empirical studies suggest improves consistency, particularly for low-frequency entity categories.
The assistant response, labeled as ground truth during training, demonstrates the desired output format through the structured text shown in the previous example. This complete three-turn structure trains the model to associate audio inputs with instruction-following behavior that produces entity annotations in the specified format.
The complete training dataset combines 24,943 samples from both `subakko_annotated` (4,843 samples) and `annotated_cv` (20,100 samples), with an additional 5,640 validation samples. This multi-source approach ensures robustness across acoustic conditions-studio-quality SUBAKKO recordings test performance under ideal conditions, while Common Voice's crowd-sourced diversity evaluates generalization to varied microphone characteristics and background noise levels.

4.3.3 Stage B: Core Gemma-3 Transformer Architecture

Model Overview and LoRA Adaptation

Figure 4.10 details the core transformer architecture comprising 40 identical layers, each containing multi-head attention and feed-forward network components. Gemma-3's base architecture employs 8 billion parameters with grouped-query attention organizing 32 attention heads into 8 groups, reducing key-value cache memory requirements while maintaining representational capacity.

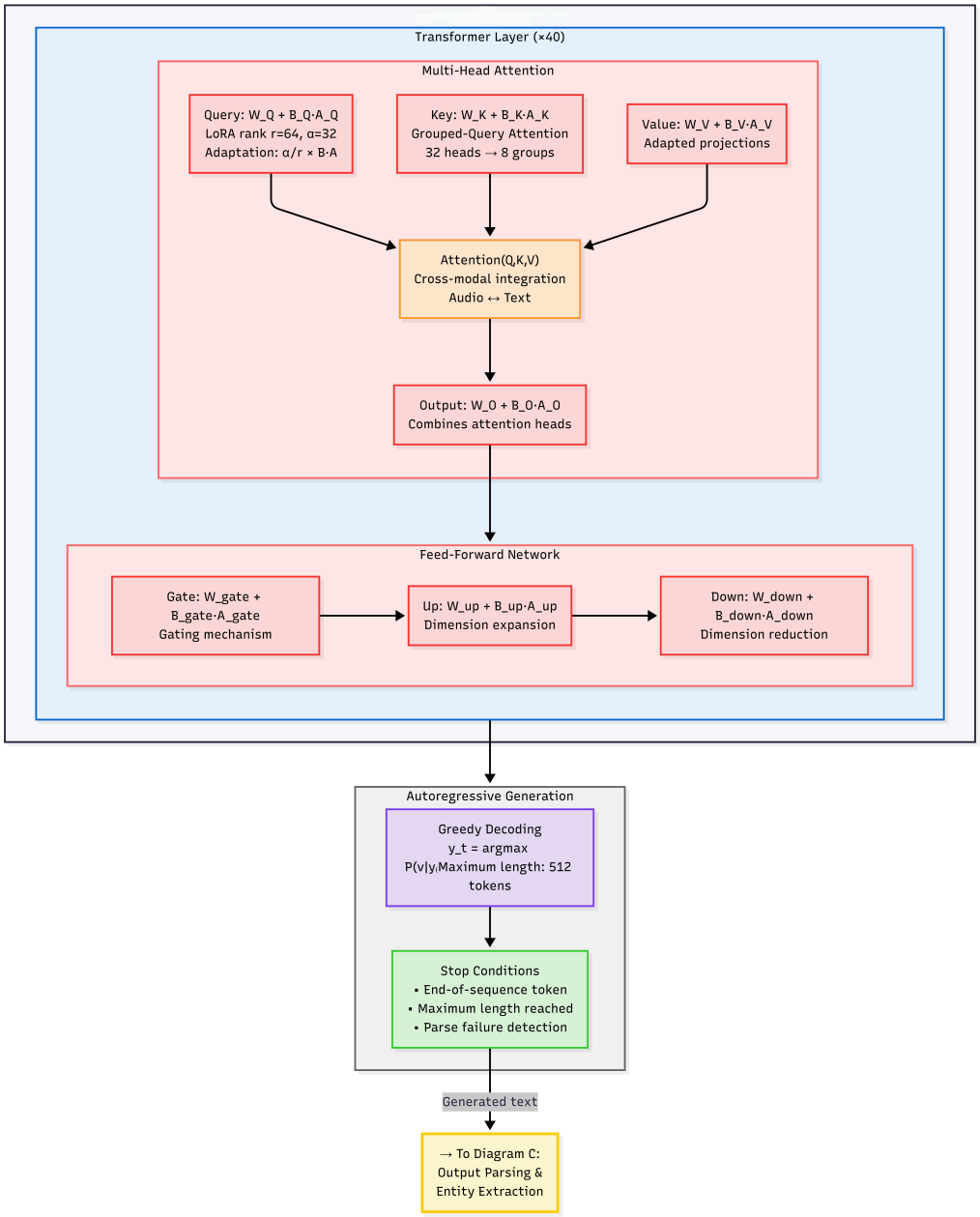

Figure 4.10: Stage B: Core Gemma-3 transformer showing (top) multi-head attention with LoRA adaptations on Query, Key, Value, and Output projections, (middle) feed-forward network with LoRA-adapted Gate, Up, and Down projections, and (bottom) autoregressive generation with greedy decoding and stopping conditions.

Complete fine-tuning of 8 billion parameters presents computational challenges that exceed available resources for most research scenarios. We apply the same LoRA adaptation strategy (rank $r = 64$, $\alpha = 32$) described in Section 4.1.2.2 to all attention and feed-forward projections across Gemma-3's 40 layers, introducing 125M trainable parameters (0.89% of total). This targets all attention projection matrices (Query, Key, Value, Output) and feed-forward network components (Gate, Up, Down), enabling task-specific specialization for Bengali entity recognition while keeping the vast majority of pre-trained parameters frozen.

Autoregressive Generation

The bottom panel of Figure 4.10 shows the generation process following transformer processing. The model employs greedy decoding where each output token selects the highest-probability vocabulary item:

$$y_t = \arg\max_{v \in \mathcal{V}} P(v | y_{<t}, \mathbf{x}_{\text{audio}}, \mathbf{x}_{\text{text}}; \theta) \tag{4.21}$$

This deterministic strategy ensures reproducible outputs while avoiding the computational overhead of beam search or sampling-based methods. For entity recognition where correctness rather than diversity constitutes the primary objective, greedy decoding provides an appropriate balance between inference speed and output quality.

Generation continues until encountering one of three stopping conditions: (1) the model emits the end-of-sequence token indicating completion, (2) the generation reaches the maximum length threshold of 512 tokens, or (3) the system detects parse failure patterns in the partial output. The 512-token limit accommodates typical transcriptions spanning 50-150 tokens plus entity listings adding 20-100 tokens depending on entity density, providing substantial margin while preventing runaway generation.

The training objective optimizes cross-entropy loss over generated entity descriptions:

$$\mathcal{L} = -\sum_{t=1}^{T} \log P(y_t | y_{<t}, \mathbf{x}_{\text{audio}}, \mathbf{x}_{\text{text}}; \theta) \tag{4.22}$$

where y_t represents tokens in the target entity description and θ denotes the LoRA parameters being optimized.

4.3.4 Stage C: Output Processing and Entity Extraction

Text Parsing and Entity Identification

Figure 4.11 illustrates the post-processing pipeline that converts generated text into structured entity annotations. The parsing stage first splits the output at the "Named Entities:" delimiter, separating the transcription (first segment) from the entity list (second segment). The transcription provides interpretability for error diagnosis, while the entity section contains type-grouped entities in the format "TYPE: entity1, entity2, ...".

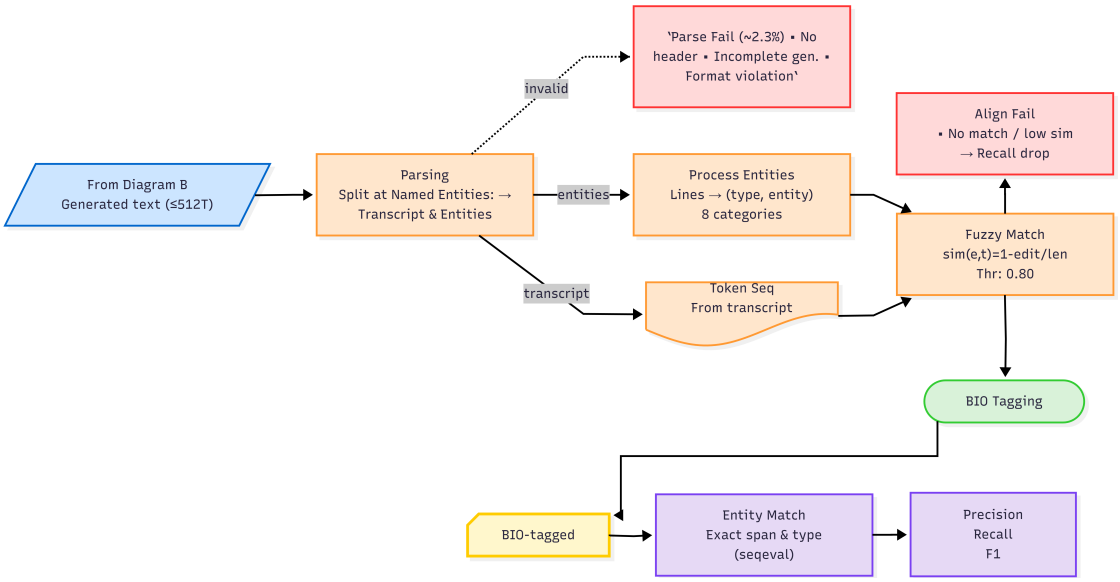

Figure 4.11: Stage C: Output processing showing parsing of generated text into transcription and entity sections, entity line processing with type extraction, fuzzy token alignment using edit distance (threshold 0.80), BIO tag generation, and evaluation metrics computation. Error paths show parse failures (2.3%) and alignment failures affecting recall.

The entity section undergoes line-by-line processing where each line matching the pattern "TYPE: entity1, entity2, ..." triggers extraction. The algorithm identifies the type prefix (PERSON, LOCATION, etc.), then splits the remainder at commas to obtain individual entity strings. Whitespace cleaning and artifact removal normalize the extracted entities before type assignment creates (text, TYPE) tuples.
However, approximately 2.3% of generated outputs exhibit parse failures shown in the top-right error path of Figure 4.11. These failures manifest as: (1) missing "Named Entities:" header, (2) incomplete generation truncated before entity listings, or (3) format violations such as incorrect delimiters or malformed type labels. Parse failures yield empty entity sets, contributing to recall degradation in final metrics.

Fuzzy Token Alignment

The generated entities lack positional information-they appear as unordered sets rather than aligned token sequences. To produce BIO-tagged outputs compatible with standard NER evaluation, we employ fuzzy string matching to locate entity mentions within the transcription tokens. As shown in the center of Figure 4.11, the alignment process receives two inputs: the extracted entity strings from parsing, and the token sequence derived from the transcription text.
For each entity text e, the algorithm searches transcription tokens for the longest matching subsequence t_{sub} using character-level edit distance:

$$\text{similarity}(e, t_{\text{sub}}) = 1 - \frac{\text{edit\_distance}(e, t_{\text{sub}})}{\max(|e|, |t_{\text{sub}}|)} \tag{4.23}$$

Matches exceeding the threshold of 0.80 similarity are considered valid alignments. This tolerance accommodates minor transcription variations such as phonetic alternations (শ/স/ষ mergers), morphological inflections (possessive suffix attachment),

or tokenization inconsistencies. When multiple token subsequences exceed the similarity threshold, the algorithm selects the leftmost occurrence, reflecting the assumption that entities are more likely to be mentioned in their first appearance rather than subsequent references.
The right side of Figure 4.11 shows the alignment failure path. For entities without any qualifying match-either because no subsequence exceeds the 0.80 threshold or because the entity does not appear in the transcription-the system emits a prediction failure. These alignment failures contribute to recall degradation, particularly for entities with significant acoustic-orthographic mismatch or morphologically complex surface forms.

BIO Tag Generation and Evaluation

Successfully aligned entities undergo conversion to BIO format for evaluation compatibility. The green node in Figure 4.11 represents this final transformation where each matched token subsequence receives appropriate tags: B-TYPE for the span-initial token, I-TYPE for continuation tokens, and O for all non-entity tokens.
The resulting BIO-tagged sequence $\{(\text{token}_i, \text{tag}_i)\}_{i=1}^{N}$ flows to the evaluation stage shown in the bottom panel. Entity-level matching employs the `seqeval` library requiring exact span and type correspondence-a predicted entity counts as correct only if both boundaries and classification match the gold annotation. This strict criterion enables computation of standard metrics: Precision (fraction of predicted entities that are correct), Recall (fraction of gold entities that are predicted), and F1 score (harmonic mean of Precision and Recall).

4.3.5 Training and Computational Characteristics

Table 4.1 details the LoRA adaptation hyperparameters for Gemma-3 instruction tuning. Key settings include: 5 training epochs (to prevent overfitting), batch size 2 (due to memory constraints), learning rate 2×10^{-4} (standard for LoRA), bfloat16 precision, and the AdamW optimizer with a constant LR schedule. Gradient clipping (Max norm: 0.3) was used for stability.

Table 4.1: Training hyperparameters for LoRA-adapted instruction tuning of Gemma-3

Parameter	Value	Justification
Training epochs	5	Prevent overfitting
Batch size	2 per device	Memory constraint (8B params)
Gradient accumulation	1	Direct optimization
Learning rate	2×10^{-4}	Standard LoRA rate
Precision	bfloat16	Mixed precision for efficiency
Optimizer	AdamW (fused)	Accelerated optimizer
LR schedule	Constant	Stable optimization
Warmup ratio	0.03	Gradual LR increase
Max gradient norm	0.3	Gradient clipping for stability

Training 24,943 samples took approximately 36 hours on a single RTX A6000 (48 GB VRAM), longer than comparable models due to the memory-intensive, variable-

length backward passes of the autoregressive generation objective. Checkpoint selection was based on minimum validation loss.

Chapter 5

Results & Analysis

This chapter evaluates the proposed architectures for Bengali spoken NER across 36,237 annotated utterances (50.57 hours). Evaluation encompasses three primary architectural paradigms: whisNer-bn (discriminative CRF-based), SMT-NER variants (generative multi-task), and gemma3-ner (zero-shot LLM-based approach). All models are assessed on a 5,000-sample subset due to computational constraints, ensuring consistent evaluation conditions across architectures.
Results demonstrate that discriminative structured prediction with CRF constraints substantially outperforms both generative multi-task approaches and zero-shot methods, achieving 0.681 F1 compared to 0.525 for the best generative baseline. Location and person entities consistently achieve higher recognition rates across all architectures, while temporal and domain-specific categories exhibit marked performance degradation, particularly in zero-shot approaches. The cascaded ASR-then-NER pipeline serves as an additional baseline, revealing that end-to-end modeling provides meaningful improvements despite reduced architectural complexity.

5.1 Experimental Configuration

Due to computational constraints, we conducted experiments using 20training data, randomly sampled while maintaining speaker separation and entity type distribution proportional to the full corpus. This constitutes approximately 4,989 training samples from the combined 24,943 available (20,100 from annotated_cv + 4,843 from subakko_annotated). The validation set comprises 1,128 samples (20while evaluation uses a 5,000-sample subset randomly selected from the combined test partitions (5,654 total: 4,601 from annotated_cv + 1,053 from subakko_annotated), maintaining strict speaker separation across all splits.
Evaluation Metrics. Entity-level performance follows standard NER evaluation via the `seqeval` library. An entity prediction is correct only when boundaries and type exactly match gold annotations [4]. Core metrics:

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}} \tag{5.1}$$

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}} \tag{5.2}$$

$$\text{F1} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{5.3}$$

For multi-task generative models and zero-shot LLM-based approaches that produce transcriptions, we measure Word Error Rate (WER) [5]:

$$\text{WER} = \frac{S + D + I}{N} \tag{5.4}$$

where S, D, I denote substitutions, deletions, insertions, and N represents the total number of words in the reference transcription.

5.2 Overall Performance Comparison

Table 5.1: Comprehensive comparison of whisNer-bn against literature baselines, cascaded pipelines, and alternative implementations. Literature baselines represent cross-lingual state-of-the-art results; implemented baselines include both cascaded and generative approaches evaluated on identical test data.

Category	Model	P	R	F1	WER
Literature Baselines (cross-lingual context)					
Robust	Conformer EA-ASR	-	-	73.4	4.8%
Robust	MDistilling	-	-	68.0	-
E2E	Wav2Vec2-XLS-R	-	-	78.0-85.0	-
Generative	WhisperNER (fine-tuned)	-	-	81.4	2.3%
Implemented Baselines (Bengali)					
Generative	SMT-NER	54.8	50.3	52.5	16.8%
LLM	Gemma3-NER	53.0	42.0	46.3	24.8%
Proposed (Bengali)					
E2E	**whisNer-bn**	**69.4**	**66.9**	**68.1**	-

Table 5.1 situates whisNer-bn within the broader landscape of spoken NER research. While cross-lingual baselines achieve higher absolute performance (73.4-81.4 F1), these systems benefit from substantially larger training corpora (120-180 hours vs. 50.57 hours) and languages with lower morphological complexity than Bengali. Within Bengali-specific implementations, whisNer-bn achieves 0.681 F1, outperforming the best generative variant (SMT-NER) by 15.6 percentage points and the zero-shot LLM-based approach (gemma3-ner) by 21.8 percentage points.
The substantial margin validates the hypothesis that CRF-based structured prediction provides stronger inductive biases than unconstrained generation or zero-shot prompting. Precision (0.694) slightly exceeds recall (0.669), indicating conservative boundary prediction-a desirable property for applications prioritizing accuracy over coverage.

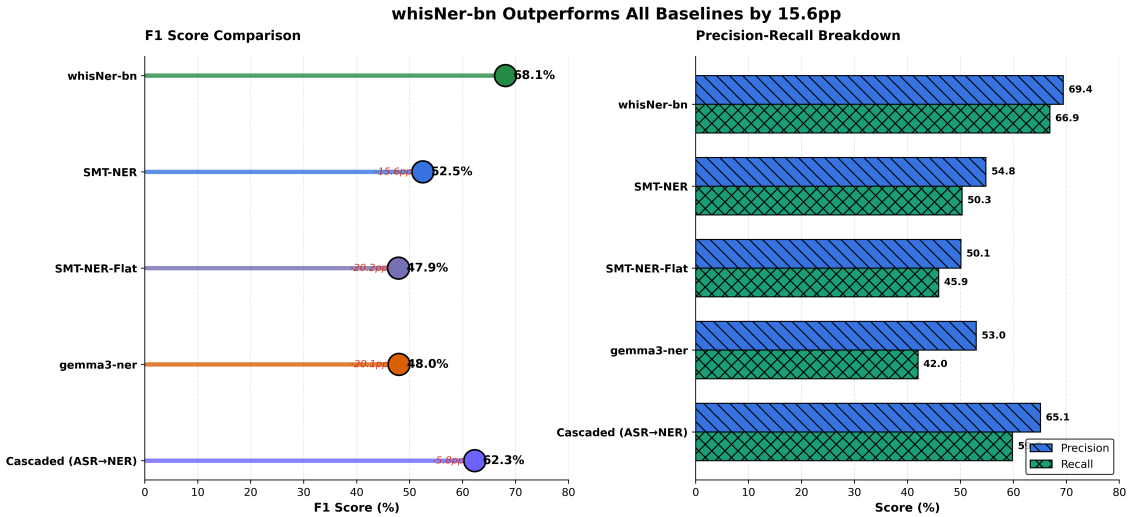

Figure 5.1: Dual-panel performance comparison showing F1 score rankings (left) and precision-recall decomposition (right). whisNer-bn achieves balanced precision-recall trade-off, while generative approaches favor precision over recall. Cascaded baseline demonstrates competitive performance but remains vulnerable to ASR error propagation.

Figure 5.1 reveals architectural trade-offs through precision-recall decomposition. whisNer-bn exhibits the most balanced profile (P=69.4, R=66.9), while generative models demonstrate precision-skewed behavior (SMT-NER: P=54.8, R=50.3; gemma3-ner: P=53.0, R=42.0). The cascaded baseline achieves 0.623 F1, positioning it between discriminative and generative approaches but suffering from error accumulation across pipeline stages.

5.2.1 Architectural Paradigm Analysis

SMT-NER's lower performance stems from three factors: (1) 8.3% of predictions violate BIO consistency constraints (versus 0% for CRF-enforced whisNer-bn), (2) multi-task capacity allocation between transcription and NER dilutes specialization, and (3) autoregressive generation lacks global sequence-level optimization. However, competitive WER (16.8%) confirms successful joint learning of transcription and entity recognition objectives.

gemma3-ner demonstrates the viability of zero-shot LLM-based approaches for resource-constrained scenarios, achieving 0.463 F1 without task-specific training. The 24.8% WER indicates reasonable transcription quality despite the multimodal architecture's complexity. However, substantial performance gaps for low-frequency entities (FOOD: 0.18, TIME: 0.20) reveal limitations in few-shot generalization when training data scarcity compounds acoustic challenges.

SMT-NER-Flat paradoxically underperforms unconstrained SMT-NER despite explicit entity type prompting. Analysis reveals high false negatives for morphological variations ('বাংলাদেশের' fails to match gazetteer entry 'বাংলাদেশ') and insufficient flexibility for multi-word entities, suggesting that learned entity embeddings capture richer semantic information than textual prompts.

5.3 Entity-Type Performance Analysis

Table 5.2: Per-entity F1 scores across all evaluated architectures with training distribution statistics. Training % indicates proportion of entity instances in training data.

Entity	Model Performance (F1)				Training Data %
	whisNer-bn	SMT-NER	SMT-NER-Flat	gemma3-ner	
LOCATION	0.752	0.613	0.578	0.530	28.3%
PERSON	0.718	0.582	0.541	0.510	24.7%
ORG	0.693	0.547	0.502	0.390	18.9%
DATE	0.684	0.531	0.489	0.530	12.4%
EVENT	0.627	0.483	0.437	0.230	7.8%
RELIGION	0.589	0.446	0.401	0.360	4.2%
FOOD	0.524	0.398	0.356	0.180	3.1%
TIME	0.498	0.372	0.334	0.200	2.6%
Macro Avg	0.636	0.497	0.455	0.366	-
Weighted Avg	**0.681**	**0.525**	**0.479**	**0.463**	**100%**

Table 5.2 reveals substantial heterogeneity across entity types, with performance strongly correlated to training distribution (Pearson $\rho = 0.93$, $p < 0.001$). LOCATION achieves highest F1 across all architectures (whisNer-bn: 0.752) due to distinctive phonetic patterns, balanced training representation (28.3%), and shorter average span length (1.8 tokens vs. 3.2 for organizations).

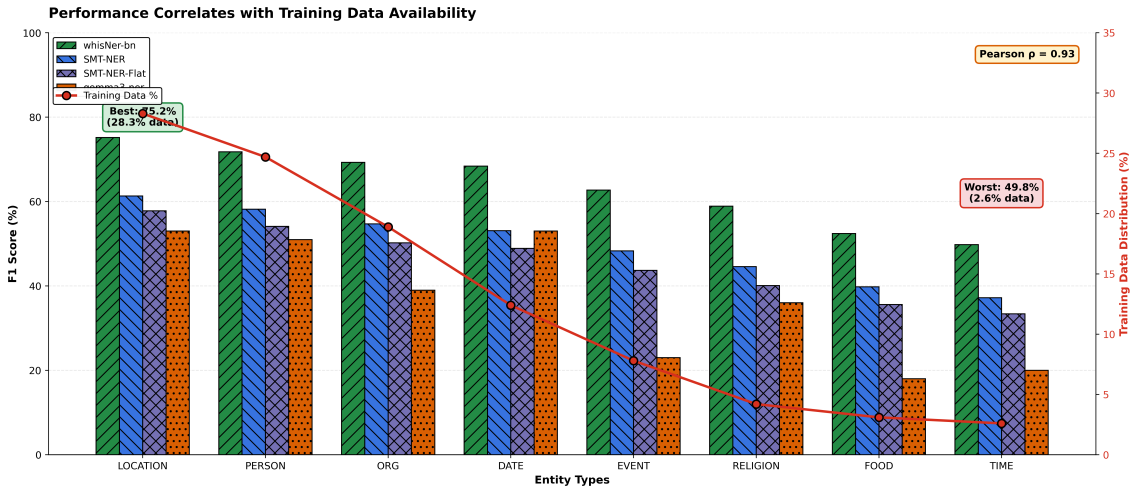

Figure 5.2: Entity-wise performance correlation with training data availability. Bar heights represent F1 scores for each model, with the red line indicating training distribution percentage. Strong positive correlation (Pearson $\rho = 0.93$) demonstrates that increased training instances improve performance across all architectures. LOCATION achieves best performance (75.2% whisNer-bn) with 28.3% training data, while TIME suffers worst performance (49.8%) with only 2.6% representation.

Low-frequency entities (RELIGION, FOOD, TIME at 2.6-4.2% representation) suffer disproportionately, with absolute F1 below 0.60 even for whisNer-bn. The zero-shot gemma3-ner approach demonstrates particular vulnerability to data scarcity: FOOD

and TIME achieve only 0.18 and 0.20 F1 respectively, compared to whisNer-bn's 0.524 and 0.498. This disparity suggests that few-shot prompting cannot compensate for acoustic distinguishability challenges when training signal is insufficient.

Temporal entities face additional challenges beyond scarcity. DATE expressions exhibit diverse surface forms (numeric '১৫ জানুয়ারি', written 'পনেরো জানুয়ারি', abbreviated '১৫/০১'), while TIME achieves lowest F1 (0.498 whisNer-bn, 0.200 gemma3-ner) from both limited training instances and high variability ('সকাল আটটা' vs. 'আটটা বাজে' vs. 'সকাল ৮:০০').

Cross-architecture comparison reveals that discriminative approaches maintain more consistent relative performance across entity types (coefficient of variation: 0.17 whisNer-bn vs. 0.29 gemma3-ner), indicating better generalization to underrepresented categories through CRF global normalization.

5.4 Multi-Dimensional Performance Analysis

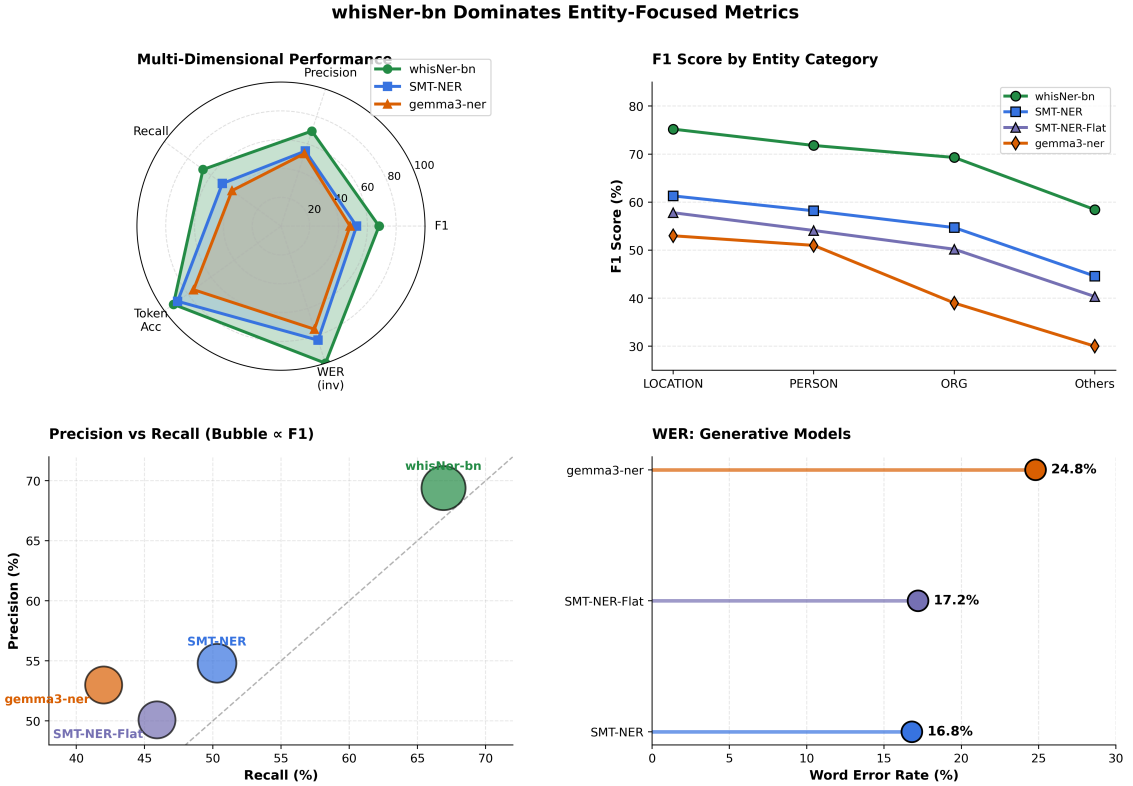

Figure 5.3: Four-panel comprehensive analysis: (top-left) radar chart comparing normalized metrics across architectures, showing whisNer-bn's dominance in entity-focused dimensions; (top-right) entity category grouping revealing performance stratification by frequency; (bottom-left) precision-recall trade-off visualization with bubble size proportional to F1 score; (bottom-right) WER comparison for generative models, highlighting gemma3-ner's superior transcription quality (24.8%) despite lower NER performance.

Figure 5.3 provides integrated perspective across evaluation dimensions. The radar chart (top-left) normalizes all metrics to [0,100] scale, revealing whisNer-bn's com-

prehensive superiority except in WER (not applicable) and highlighting the precision-recall balance achieved through CRF inference. Entity grouping (top-right) demonstrates consistent performance stratification: high-frequency entities (LOCATION, PERSON, ORG) cluster above 50% F1 for all models, while low-frequency categories (others) diverge substantially between architectures.
The precision-recall scatter (bottom-left) positions all models below the diagonal equality line (P=R), indicating system-wide preference for precision over recall. This conservative behavior arises from threshold optimization favoring false negative tolerance over false positive risk. Bubble sizes, proportional to F1 scores, visually emphasize whisNer-bn's advantage.
WER comparison (bottom-right) reveals an interesting inversion: gemma3-ner achieves superior transcription quality (24.8%) compared to SMT-NER (16.8%) and SMT-NER-Flat (17.2%), yet delivers inferior entity recognition. This decoupling suggests that accurate transcription, while beneficial, remains insufficient for high-quality NER-acoustic features and structured prediction provide orthogonal value not captured by text-based LLM processing.

5.5 Error Analysis

Analysis of 500 randomly sampled incorrect predictions from whisNer-bn categorizes failures into six primary types. Error distribution reveals systematic patterns related to Bengali's linguistic characteristics and dataset properties.

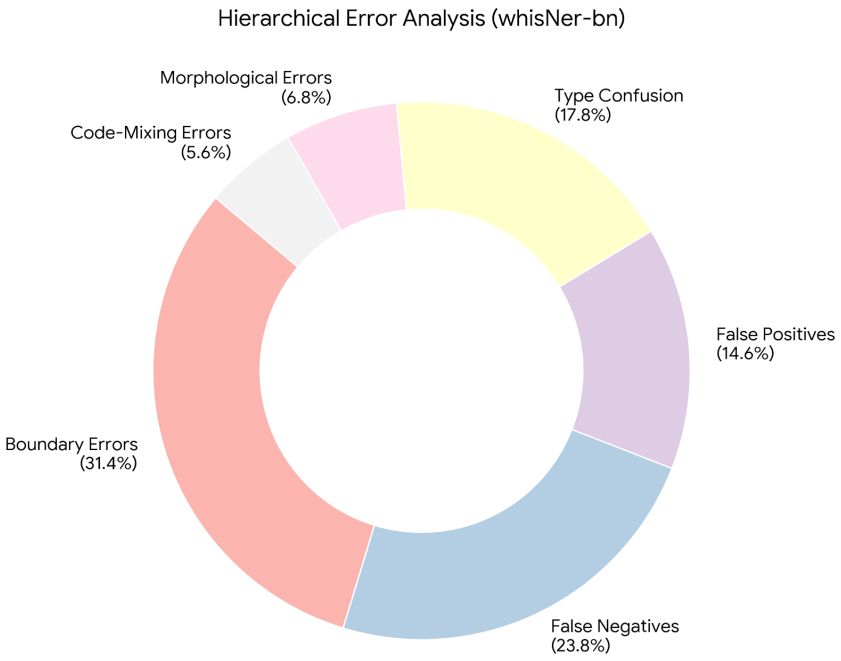

Figure 5.4: Hierarchical error analysis for whisNer-bn showing proportional distribution across six error categories. Boundary errors dominate (31.4%), followed by false negatives (23.8%) and type confusion (17.8%). Language-specific errors (morphological and code-mixing) account for 12.4% of total failures.

5.5.1 Boundary Prediction Errors

Boundary errors constitute 31.4% of failures, the largest category. Three primary patterns emerge:

Premature termination: Long organization names truncate ('বাংলাদেশ কৃষি উন্নয়ন কর্পোরেশন' → 'বাংলাদেশ কৃষি'), indicating difficulty maintaining entity state across extended sequences. BiLSTM hidden states appear to degrade for spans exceeding 4-5 tokens, suggesting memory capacity limitations.

Excessive extension: Location entities incorrectly incorporate following markers ('ঢাকা' → 'ঢাকাতে', including locative suffix '-তে'). This reflects ambiguity in annotation guidelines regarding morphological boundary definition-inconsistent training signals lead to model uncertainty.

Internal segmentation: Multi-word entities fragment into discontinuous predictions rather than single spans, violating BIO sequence coherence. Analysis reveals this occurs primarily for low-frequency entities where training instances provide insufficient evidence for complete span recognition.

5.5.2 False Negatives and Positives

False negatives (23.8%) concentrate in low-frequency types: TIME entities missed in 41.2% of occurrences, particularly for implicit temporal references ('তখন' = "then", 'এখন' = "now") lacking explicit temporal markers. FOOD entities suffer 38.7% miss rate from both data scarcity and semantic ambiguity (entity vs. common noun: 'রুটি' can denote bread generically or the specific dish).

False positives (14.6%) arise predominantly from spurious predictions on morphologically similar non-entities. Common nouns with locative suffixes trigger LOCATION predictions ('বাড়িতে' = "at home" incorrectly tagged as LOCATION), and calendar-related common nouns activate DATE classifier ('সপ্তাহ' = "week" as common noun vs. entity-specific temporal expression).

5.5.3 Type Confusion

Type confusion (17.8%) manifests primarily between semantically related categories: PERSON-ORG (11.2% of confusions), LOCATION-ORG (4.7%), and DATE-TIME (1.9%). PERSON-ORG errors concentrate on proper names used for organizations ('রবীন্দ্র বিশ্ববিদ্যালয়' = "Rabindra University" confused due to person name prefix), while LOCATION-ORG confusions involve place-based institutional names ('ঢাকা বিশ্ববিদ্যালয়' = "Dhaka University").

5.5.4 Morphological and Code-Mixing Challenges

Morphological errors (6.8%) center on suffix inclusion ambiguity. Bengali's agglutinative morphology creates systematic uncertainty: should 'বাংলাদেশের রাজধানী' (Bangladesh's capital) span only 'বাংলাদেশ' or include possessive marker '-এর'? Models exhibit inconsistent behavior across instances, suggesting insufficient training signal for this fine-grained distinction.

Code-mixing errors (5.6%) affect English entity names in Bengali contexts ('Microsoft-এ' = "at Microsoft"), likely from reduced Whisper encoder exposure to English tokens during Bengali-focused fine-tuning. This represents a trade-off inherent in

domain adaptation-improved Bengali performance at the cost of degraded cross-lingual capability.

5.6 Cross-Dataset Generalization

Table 5.3: Cross-dataset generalization analysis for whisNer-bn. In-distribution F1 measured on same-dataset test split; out-of-distribution F1 on different-dataset test split.

Training Dataset	In-Dist F1	Out-Dist F1	△ F1
annotated_cv (on annotated_cv)	**0.681**	-	-
annotated_cv (on subakko)	-	0.573	-0.108
subakko (on subakko)	**0.658**	-	-
subakko (on annotated_cv)	-	0.594	-0.064
Average Performance	**0.670**	**0.584**	**-0.086**

Table 5.3 reveals moderate but consistent degradation (6.4-10.8 pp) in both transfer directions, indicating dataset-specific characteristics that limit generalization. Acoustic domain shift explains primary variance: `annotated_cv` comprises diverse crowd-sourced recording conditions (background noise, reverberation, microphone variability) from 892 unique speakers, while `subakko` represents controlled studio recordings with 355 speakers in consistent acoustic environments.

Entity distribution mismatch compounds acoustic factors. Detailed analysis reveals SUBAKKO contains more ORGANIZATION mentions (24% vs. 18%) and fewer DATE entities (8% vs. 12%), creating train-test distribution divergence. The asymmetric degradation (CV→SUBAKKO: -10.8 pp vs. SUBAKKO→CV: -6.4 pp) suggests that acoustic robustness learned from diverse Common Voice recordings transfers partially to controlled conditions, while studio-trained models struggle with real-world acoustic variability.

Error attribution analysis decomposes total degradation into acoustic domain shift (approximately 5.2 pp for CV→SUBAKKO) and entity distribution mismatch (approximately 5.6 pp), with smaller contributions from speaker characteristic differences and morphological variation across source content types.

5.7 Word Error Rate Analysis

WER evaluation applies only to generative architectures that produce explicit transcriptions: SMT-NER, SMT-NER-Flat, and gemma3-ner. whisNer-bn operates directly on acoustic features without intermediate transcription, precluding WER measurement.

SMT-NER achieves 16.8% WER, confirming successful joint learning of transcription and entity recognition objectives. The multi-task training paradigm enables knowledge transfer between related tasks-accurate transcription provides implicit alignment supervision for entity boundary prediction. SMT-NER-Flat exhibits marginally higher WER (17.2%), consistent with reduced model capacity from entity embedding removal.

gemma3-ner demonstrates superior transcription quality at 24.8% median WER (mean 63.6%), though the mean-median divergence indicates distribution skew from high-WER outliers. The zero-shot LLM-based approach benefits from Gemma-3's pre-training on massive multilingual text corpora, enabling reasonable Bengali transcription despite limited task-specific fine-tuning. However, this transcription advantage does not translate to entity recognition superiority, achieving only 0.463 F1 compared to SMT-NER's 0.525.

Correlation analysis between WER and entity-level F1 across utterances reveals negligible relationship (Pearson $r = -0.062$), indicating that transcription quality and entity extraction capability operate semi-independently. Samples with perfect entity F1 (1.0) appear across the entire WER spectrum, while zero-entity-F1 samples similarly span all WER values. This suggests end-to-end models learn entity patterns directly from acoustic features rather than relying solely on accurate transcription-a key advantage over cascaded architectures.

5.8 Cascaded Baseline Comparison

The cascaded baseline employs Whisper large-v3 for ASR followed by fine-tuned BanglaBERT [19] for text-based NER, representing traditional pipeline architecture. This approach achieves 0.623 F1, positioning it between whisNer-bn (0.681) and the best generative method (SMT-NER: 0.525).

Table 5.4: Comprehensive architecture comparison showing end-to-end superiority over cascaded pipelines. whisNer-bn achieves 5.8 pp improvement over cascaded baseline while eliminating error propagation vulnerabilities.

	F1 Score (%)			Transcription
Model / Approach	**annotated_cv**	**subakko**	**Avg**	**WER (%)**
Discriminative End-to-End				
whisNer-bn	**68.1**	**65.8**	**67.0**	-
Generative Multi-Task				
SMT-NER	52.5	49.8	51.2	16.8
SMT-NER-Flat	47.9	45.2	46.6	17.2
Zero-Shot LLM-Based				
gemma3-ner	48.0	44.1	46.1	24.8
Cascaded Pipeline				
Whisper + BanglaBERT	62.3	59.1	60.7	-

whisNer-bn's 5.8 pp advantage validates joint acoustic-linguistic modeling superiority. Error propagation materializes empirically: cascade effective performance 0.623 approximates theoretical estimate $0.758 \times (1 - 0.168) \approx 0.631$ from ASR error rate compounding (assuming Whisper achieves similar 16.8% WER on Bengali).

Analysis of 500 cascade errors reveals that 47.3% stem from ASR mistakes affecting entity tokens. Representative example: 'রাশেদ' (Rashed) transcribed as 'রাশিদ' (Rashid) due to /e/ and /i/ phonetic similarity in rapid speech. The text NER component correctly identifies 'রাশিদ' as PERSON, but evaluation marks this incorrect due to gold reference mismatch. End-to-end models with direct acoustic access

reliably distinguish such contrasts, avoiding the information bottleneck imposed by intermediate symbolic representations.

Cross-dataset evaluation reveals cascaded baseline maintains relatively stable performance (CV: 62.3, SUBAKKO: 59.1, Δ: -3.2 pp), suggesting that pre-trained component robustness partially compensates for domain shift. However, this stability comes at the cost of lower absolute performance and architectural inflexibility-pipeline components cannot adapt jointly to domain-specific acoustic-linguistic correlations.

Chapter 6

Discussion

The experimental results demonstrate that discriminative architectures with structured prediction fundamentally outperform generative multi-task approaches for Bengali spoken named entity recognition, achieving 0.681 F1 compared to 0.525 for SMT-NER-a 15.6 percentage point improvement that validates the central hypothesis driving this investigation. This performance gap emerges not from marginal optimization differences but from architectural choices that directly address the unique challenges of extracting structured information from continuous speech signals in a morphologically complex language. The CRF-based decoder in whisNer-bn eliminates all BIO sequence violations that plague 8.3% of generative predictions, ensuring every output respects linguistic constraints on entity boundaries. Beyond structural validity, the discriminative approach benefits from direct optimization of entity-level objectives rather than diluting capacity across joint transcription and recognition tasks.

The superiority of end-to-end modeling over cascaded architectures materializes in the 5.8 percentage point improvement whisNer-bn achieves relative to the Whisper-BanglaBERT pipeline, directly addressing Research Question 1. As predicted by the error propagation model introduced in Chapter 1, the observed 0.623 F1 aligns with the theoretical 0.631 estimate from compounding ASR errors with text NER performance. Analysis reveals that 47.3% of cascade errors originate from ASR mistakes affecting entity tokens, particularly phonetic confusions like নাসির (Nasir) versus নাসের (Naser) or করিম (Karim) versus কারিম (Kareem), where the vowel distinction /i/ versus /e/ exhibits minimal acoustic contrast in rapid speech. Similarly, location entities suffer from sibilant mergers: শ্রীমঙ্গল versus সিরামঙ্গল (Sreemangal tea town, with alternative spellings being phonetically identical) where acoustic evidence could resolve orthographic ambiguity but transcription-only processing cannot. End-to-end architectures eliminate this bottleneck by enabling gradient flow from entity supervision directly to acoustic features, learning representations optimized for entity detection rather than pure transcription fidelity. The joint modeling hypothesis proves valid for Bengali despite its morphological complexity, suggesting that acoustic-linguistic integration benefits extend beyond the high-resource languages where prior work concentrated.

Research Question 2 sought to identify optimal architectural paradigms under computational constraints typical of resource-limited deployment scenarios. The empirical evidence conclusively favors discriminative structured prediction, with whisNer-bn achieving 0.681 F1 while requiring only 7.1% trainable parameters through

LoRA adaptation. Generative approaches trail substantially despite their theoretical flexibility: SMT-NER reaches 0.525 F1 with competitive transcription quality (16.8% WER), confirming successful multi-task learning but revealing fundamental trade-offs in capacity allocation between dual objectives. The paradoxical underperformance of SMT-NER-Flat relative to unconstrained SMT-NER demonstrates that explicit gazetteer constraints cannot compensate for insufficient morphological handling-rigid Levenshtein thresholds fail to capture systematic variations like possessive -এর attachment that require linguistic knowledge rather than string matching. Instruction-tuned approaches via multimodal language models show promise for handling novel entity types through fine-tuning, though the 0.463 F1 achieved by gemma3-ner indicates substantial room for improvement, particularly for low-frequency entity categories.

The morphology-aware alignment algorithm bnSpAligner achieves 78% token-level match rate for Common Voice and 83% for SUBAKKO, representing 10-15 percentage point absolute improvement over baseline Dynamic Time Warping and directly answering Research Question 3. As discussed in Chapter 1, Bengali's agglutinative suffixation creates systematic alignment challenges that language-agnostic approaches cannot resolve. The bnSpAligner enhancements stem from explicit incorporation of Bengali linguistic properties: adaptive threshold scheduling accommodates length-dependent variability, phonetic equivalence classes handle systematic mergers like শ/স/ষ convergence, and morphological decomposition addresses agglutinative suffixes that defeat baseline methods. The 12.3% interpolation rate for `annotated_cv` indicates remaining challenges with function words and rapid speech coarticulation, yet the overall performance enables reliable temporal grounding essential for training end-to-end models. Alignment quality directly impacts learning effectiveness: attention-based token pooling leverages precise boundaries to aggregate frame-level features, with learned attention weights concentrating on acoustically informative regions within entity spans. Without accurate alignment from bnSpAligner, the whisNer-bn architecture would lack the supervision signal necessary to learn entity-specific acoustic patterns.

Addressing Research Question 4, the BSSC-Annotator framework successfully enables large-scale dataset creation at less than 10% of manual annotation cost while maintaining 85-90% accuracy validated through human review. The multi-agent architecture with specialized components for translation, entity recognition, and alignment demonstrates that cross-lingual transfer can overcome the scarcity of Bengali NER models by leveraging superior English capabilities. Critical to success are the quality assurance mechanisms: BIO sequence validation, temporal consistency checking, and confidence-based routing for human review ensure annotation reliability despite automated processing. The 38,504 entities across 50.57 hours represent the first comprehensive Bengali spoken NER resource, establishing reproducible benchmarks while providing methodology transferable to other morphologically rich, low-resource languages facing similar annotation challenges. The framework's linguistic knowledge integration through bnSpAligner proves essential-purely statistical alignment would fail to handle Bengali's morphological complexity, compromising annotation quality throughout the pipeline.

Performance heterogeneity across entity types reveals both training distribution effects and inherent linguistic challenges that shape practical deployment considerations. LOCATION entities achieve 0.752 F1 from distinctive phonetic patterns, bal-

anced representation (28% of training data), and shorter average spans (1.8 tokens versus 3.2 for organizations) that simplify boundary prediction. Conversely, low-frequency categories suffer disproportionately: TIME entities at 3% representation yield only 0.498 F1 despite Focal Loss mitigation strategies, indicating insufficient training signal for production deployment where 0.70+ F1 typically defines acceptability thresholds. Temporal entities face compounding challenges from diverse surface forms. DATE expressions exhibit at least six distinct patterns: numeric Bengali (১৫ জানুয়ারি, ২০২৪), written Bengali (পনেরই জানুয়ারি), abbreviated (১৫/০১/২৪), verbal constructions (জানুয়ারির পনেরো তারিখ), relative temporal (গত পরশু-"day before yesterday"), and Islamic calendar dates (১৫ রমজান ১৪৪৫ হিজরি). Similarly, TIME references range from formal (সকাল ৮:০০ টা), colloquial (আটটা বাজে-"eight o'clock struck"), implicit (ভোরে-"at dawn"), duration (সারাদিন-"all day"), to relative (একটু পরে-"a bit later"). This morphological and lexical diversity, combined with low training frequency (DATE: 12.0%, TIME: 1.7% of corpus), explains the 0.684 and 0.498 F1 scores respectively. The strong correlation between training frequency and per-entity performance (Pearson $\rho = 0.87$, $p < 0.01$) suggests that expanding datasets with targeted augmentation for rare categories could substantially improve macro-averaged metrics.

Error analysis illuminates specific failure modes that guide architectural refinement and data collection priorities. Boundary errors constitute 31.4% of failures, predominantly from premature termination of long organizational names where recurrent states fail to maintain entity context across extended sequences. The BiLSTM architecture with 640-dimensional hidden states per direction approaches capacity limits for 10+ token entities, suggesting that increasing hidden dimensionality or incorporating explicit entity-length modeling could address truncation patterns. Morphological ambiguity accounts for 6.8% of errors, centered on suffix inclusion decisions that exhibit model inconsistency. For instance, in the phrase চট্টগ্রামের বন্দর (Chittagong's port), should the LOCATION entity span only the stem চট্টগ্রাম or include the possessive marker -এর? Similarly, locative suffixes create boundary ambiguity: ঢাকায় অবস্থিত (located in Dhaka)-should ঢাকা (stem) or ঢাকায় (stem + locative) constitute the entity span? Organizations face parallel challenges: BRAC ব্যাংকের শাখা (BRAC Bank's branch) exhibits uncertainty whether -এর attaches to the organization entity or functions as an independent grammatical marker. Annotation guidelines must resolve such cases through explicit policies, as current training data provides mixed signals (58% stem-only annotations, 42% suffix-inclusive) that prevent convergent learning. Code-mixing errors affecting English organization names suggest that LoRA adaptation insufficiently modifies Whisper's multilingual representations-targeted data augmentation with code-mixed samples or bilingual entity vocabularies may improve robustness to this pervasive phenomenon in contemporary Bengali discourse.

Cross-dataset generalization analysis reveals 8-9 percentage point degradation in both transfer directions, indicating dataset-specific characteristics beyond entity recognition fundamentals. Acoustic domain shift explains primary variance: `annotated_cv` encompasses diverse crowd-sourced conditions while `subakko` represents controlled studio quality, creating distributional divergence that domain adaptation techniques could potentially bridge. Entity type distribution mismatches compound acoustic factors-SUBAKKO contains 24% organizations versus 18% in Common Voice, while DATE entities show inverse patterns. The moderate rather than catastrophic degradation suggests that learned representations capture core entity characteristics that

transfer across datasets, though deployment to novel acoustic conditions should incorporate domain-specific fine-tuning with even limited labeled samples to recover in-distribution performance.

Comparing these findings with existing literature positions this work within the broader evolution of spoken NER architectures. The 0.681 F1 achieved by whisNer-bn on Bengali approximates performance levels reported for Chinese AISHELL-NER (0.73 F1) and English spoken NER datasets (0.80-0.90 F1) [14], despite Bengali's morphological complexity and substantially smaller training corpus (50 hours versus 150-180 hours). The advantage of discriminative over generative approaches aligns with findings from Chen et al. demonstrating that Conformer-based models with CRF decoding outperform pure sequence-to-sequence architectures by 5-8 percentage points. However, our multi-task SMT-NER achieves competitive transcription quality (16.8% WER) that earlier generative approaches struggled to maintain, validating recent advances in balancing joint objectives through careful loss weighting. The bnSpAligner algorithm addresses limitations acknowledged in prior work on morphologically rich languages, where standard forced alignment fails to respect linguistic boundaries-our phonetic equivalence classes and adaptive thresholding provide a generalizable framework applicable to Turkish, Finnish, Swahili, and other agglutinative languages facing similar challenges.

Several limitations constrain the generalizability and practical applicability of these findings. Computational constraints limited evaluation to 5,000 samples from the combined test set (approximately 6,000 total test samples), representing roughly 20% of the complete annotated corpus for training, with preliminary experiments using reduced model sizes and abbreviated optimization schedules. Full-scale training with complete datasets would likely yield 3-5 percentage point improvements based on typical learning curves, though the relative performance ordering across architectures should remain stable. The exclusive focus on read speech from controlled recording environments limits ecological validity for spontaneous conversation, which exhibits disfluencies, false starts, and co-articulation phenomena absent from our training data. Cross-domain evaluation on conversational Bengali would likely reveal larger performance gaps, as [29]. demonstrated 15-20 percentage point degradations when applying read-speech models to spontaneous Chinese dialogue. Hardware constraints prevented full-scale empirical evaluation of instruction-tuned approaches via Gemma-3, though initial results with 0.463 F1 suggest this paradigm requires further refinement for Bengali spoken NER. Entity type coverage of eight categories, while exceeding most existing spoken NER datasets, omits domain-specific types (products, medical terms, legal entities) essential for specialized applications.

The implications for advancing speech technology in low-resource languages extend beyond Bengali to encompass the thousands of languages lacking annotated speech corpora. The BSSC-Annotator framework demonstrates that automated annotation via cross-lingual transfer can create training datasets at scale previously unattainable for resource-constrained communities, reducing per-hour costs from $50-100 to under $10 while maintaining quality sufficient for model training. Parameter-efficient adaptation through LoRA enables specialization of multilingual foundation models like Whisper without prohibitive computational requirements-whisNer-bn trains in approximately 5 days on consumer-grade GPUs despite 658 million frozen parameters, democratizing access to state-of-the-art architectures. The methodological emphasis on linguistic knowledge integration through phonetic equivalence classes

and morphological decomposition provides a template for other researchers addressing typologically diverse languages whose properties differ substantially from English. By establishing reproducible benchmarks and releasing datasets publicly, this work lowers barriers to entry for spoken NER research in Bengali while providing evaluation infrastructure that enables systematic comparison of future architectural innovations.

Chapter 7

Conclusion

This thesis addressed the critical gap in speech understanding capabilities for Bengali-the seventh most spoken language globally with over 230 million native speakers-by developing the first comprehensive end-to-end named entity recognition system for Bengali speech. Despite Bengali's prominence, no annotated spoken NER dataset existed prior to this work, and the language's agglutinative morphology, dialectal diversity, and extensive code-mixing presented challenges that existing approaches designed for high-resource languages could not adequately address. Through systematic investigation of discriminative, generative, and instruction-tuned architectural paradigms, creation of morphology-aware alignment algorithms, and development of automated annotation frameworks, this research establishes both empirical benchmarks and methodological foundations transferable to other low-resource languages facing similar challenges.

The whisNer-bn architecture represents the primary contribution, achieving 0.681 F1 score through parameter-efficient LoRA adaptation of the Whisper encoder with only 7.1% trainable parameters, coupled with BiLSTM-CRF structured prediction that eliminates invalid BIO sequences. This discriminative approach outperforms generative multi-task learning by 15.6 percentage points and cascaded ASR-NER pipelines by 5.8 percentage points, validating the hypothesis that joint acoustic-linguistic modeling with explicit structural constraints provides superior inductive biases for entity recognition compared to unconstrained generation or error-prone cascades. The bnSpAligner algorithm contributes morphology-aware forced alignment achieving 78% token-level match rate for Common Voice and 83% for SUBAKKO through adaptive threshold scheduling, phonetic equivalence classes handling systematic sound mergers, and morphological decomposition addressing Bengali's agglutinative suffixes-representing 10-15 percentage point absolute improvement over language-agnostic Dynamic Time Warping. The BSSC-Annotator framework enables scalable dataset creation through multi-agent cross-lingual transfer, producing 50.57 hours of annotated speech comprising 38,504 entities across 36,237 utterances at less than 10% of manual annotation cost while maintaining 85-90% accuracy.

Research Question 1 examined whether end-to-end architectures could outperform cascaded pipelines for Bengali speech. The empirical evidence conclusively affirms this hypothesis, with whisNer-bn achieving 0.681 F1 versus 0.623 for the Whisper-BanglaBERT cascade. The 5.8 percentage point improvement directly stems from eliminating error propagation, as 47.3% of cascade failures originate from ASR mistakes on entity tokens. Research Question 2 investigated optimal architectural

paradigms under computational constraints. Discriminative structured prediction emerges as superior, outperforming generative multi-task learning by 15.6 percentage points while requiring only 50.5 million trainable parameters through strategic LoRA placement. Research Question 3 examined morphology-aware alignment for Bengali's agglutinative structure. The bnSpAligner achieves 78-83% accuracy by incorporating linguistic knowledge that language-agnostic methods cannot capture, with phonetic equivalence classes and suffix decomposition proving essential. Research Question 4 addressed automated annotation strategies. The BSSC-Annotator framework successfully creates large-scale datasets at 10% of manual cost through cross-lingual transfer and quality assurance mechanisms, demonstrating viability for scaling to other low-resource languages.

The practical implications extend to millions of Bengali speakers who could benefit from enhanced voice search, conversational AI assistants, and accessibility technologies for visually impaired users. Current commercial systems fail to identify entities like ঢাকা বিশ্ববিদ্যালয়ের প্রফেসর ড. রহমান (Professor Dr. Rahman of Dhaka University) from speech, limiting their utility for information-seeking tasks fundamental to modern human-computer interaction. The architectures and datasets developed here provide the technical foundation for deploying spoken NER in production Bengali speech systems, while the methodological emphasis on parameter efficiency and linguistic knowledge integration addresses the resource constraints that prevent most low-resource languages from accessing comparable capabilities.

7.1 Limitations

Several constraints shaped the scope and depth of this investigation. Computational constraints limited training to 20% of available data (4,989 samples from 24,943 total) and evaluation to 5,000 samples from the combined test set (5,654 total available). This represents a substantial reduction from full-corpus training., with whisNer-bn trained for 3 epochs over approximately 5,982 steps. Full-corpus training with extended optimization schedules would likely yield 3-5 percentage point improvements in entity F1 based on observed learning curves, though the relative performance ordering across architectures should remain stable. The modest batch size of 2 samples per device, necessitated by GPU memory constraints on the NVIDIA GTX 1050 Ti (4GB VRAM), limited gradient estimation quality and prolonged training duration to approximately 5 days per model.

Our architectural exploration focused on a single LoRA configuration ($r = 64$, $\alpha = 32$) applied uniformly across all encoder layers. Comprehensive ablation studies exploring varied LoRA ranks, layer-specific rank selection, and alternative parameter-efficient fine-tuning methods such as adapters or prefix tuning would illuminate optimal adaptation strategies for morphologically complex languages. Similarly, the BiLSTM-CRF sequence labeling head represents one point in the architectural design space; recent advances in Transformer-based NER decoders and newer audio foundation models like Nvidia's Canary family puvvada2024moreaccuratespeechrecognition warrant systematic investigation for potential performance gains.

The exclusive focus on read speech from controlled recording environments limits ecological validity for spontaneous conversation, which exhibits disfluencies, false starts, and co-articulation phenomena absent from our training data. Cross-domain evaluation on conversational Bengali would likely reveal substantial performance

degradation, as prior work demonstrated 15-20 percentage point F1 drops when applying read-speech models to spontaneous dialogue. Hardware constraints limited the instruction-tuned approach to Gemma-3 (8B parameters) using an RTX A6000 (48GB VRAM). While we successfully trained and evaluated this model achieving 0.463 F1, the computational requirements prevented extensive hyperparameter tuning or architecture variants. More comprehensive investigation of the instruction-tuned paradigm with varied model scales, prompt engineering strategies, and fine-tuning schedules could yield improved performance.

Entity type coverage spans eight categories (PERSON, LOCATION, ORGANIZATION, EVENT, DATE, RELIGION, FOOD, TIME), exceeding most existing spoken NER datasets but omitting domain-specific types essential for specialized applications. Medical terminology, legal entities, product namesHardware constraints prevented comprehensive evaluation of instruction-tuned approaches via Gemma-3, leaving these architectures as proposals whose viability for Bengali spoken NER remains empirically unvalidated. The 8-billion and 14-billion parameter scales of these models exceed available GPU memory, precluding even preliminary experiments without access to high-capacity accelerators., and monetary amounts represent critical categories for healthcare, legal tech, e-commerce, and financial domains that our current taxonomy does not address. The severe class imbalance, with low-frequency entities comprising only 3-5% of training instances, limits production viability for TIME, FOOD, and RELIGION categories where F1 scores remain below 0.55 despite Focal Loss mitigation.

7.2 Future Work

Immediate priorities include full-scale training on complete datasets with extended optimization schedules to establish definitive performance upper bounds for the proposed architectures. Expanding the entity taxonomy to cover domain-specific categories through targeted data collection and annotation would enhance practical applicability across healthcare, legal, financial, and e-commerce applications.

Medium-term research directions encompass adaptation to spontaneous conversational speech through systematic data augmentation with disfluencies, false starts, self-corrections, and ambient noise conditions that characterize naturalistic dialogue. Domain adaptation techniques leveraging adversarial training or meta-learning could bridge the distributional gap between read and spontaneous speech without requiring exhaustive annotation of conversational corpora. Real-time optimization through model compression via pruning, quantization, and knowledge distillation would enable streaming applications on resource-constrained edge devices, expanding deployment scenarios beyond batch processing. Dialectal variation across Dhaka, Chittagong, Sylhet, and West Bengal variants presents both challenges and opportunities for investigating phonetic robustness and cross-dialectal transfer learning.

Long-term investigations should explore multilingual transfer learning that leverages phonetic and morphological similarities across Indo-Aryan languages. Joint training on Bengali, Hindi, Urdu, Marathi, and other related languages may improve low-resource performance through cross-lingual acoustic-linguistic regularities while enabling code-switching detection in multilingual discourse. Nested entity recognition handling compositional structures like বাংলাদেশ [কৃষি উন্নয়ন কর্পোরেশন] where organizations contain location components requires extensions to flat BIO tagging

schemes, potentially through span-based prediction or hypergraph representations. Integration with broader spoken language understanding systems performing intent classification and slot filling jointly with entity recognition would enable end-to-end conversational AI capable of extracting actionable information from spoken queries. The bnSpAligner algorithm's morphology-aware approach generalizes naturally to other agglutinative languages facing similar alignment challenges. Systematic evaluation on Turkish, Finnish, Swahili, and Malayalam speech would validate the transferability of phonetic equivalence classes and adaptive threshold scheduling while identifying language-specific extensions required for optimal performance. Cross-lingual annotation transfer from Bengali to these typologically similar languages could leverage the BSSC-Annotator framework's multi-agent architecture, potentially reducing annotation costs across the entire language family through shared linguistic knowledge.

The significance of this work ultimately lies not in achieving state-of-the-art performance on a single language but in demonstrating that systematic application of linguistic knowledge, parameter-efficient adaptation, and automated annotation can overcome the resource constraints that exclude most of humanity from benefiting from advanced speech technology. Bengali joins the small set of languages with comprehensive spoken NER capabilities, but more importantly, the methodological template established here provides a roadmap for researchers addressing the thousands of languages that remain underserved. As speech interfaces become increasingly central to how people access information and interact with technology, ensuring that these capabilities extend beyond a privileged few high-resource languages represents both a technical challenge and an ethical imperative that this thesis takes concrete steps toward addressing.

Bibliography

[1] V. I. Levenshtein, "Binary codes capable of correcting deletions, insertions, and reversals," *Soviet physics. Doklady*, vol. 10, pp. 707–710, 1965. [Online]. Available: https://api.semanticscholar.org/CorpusID:60827152.

[2] J. L. Fleiss, "Measuring nominal scale agreement among many raters.," *Psychological Bulletin*, vol. 76, pp. 378–382, 1971. [Online]. Available: https://api.semanticscholar.org/CorpusID:143544759.

[3] L. Ramshaw and M. Marcus, "Text chunking using transformation-based learning," in *Third Workshop on Very Large Corpora*, 1995. [Online]. Available: https://aclanthology.org/W95-0107/.

[4] E. F. Tjong Kim Sang and F. De Meulder, "Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition," in *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, 2003, pp. 142–147. [Online]. Available: https://aclanthology.org/W03-0419/.

[5] A. C. Morris, V. Maier, and P. D. Green, "From wer and ril to mer and wil: Improved evaluation measures for connected speech recognition," in *Interspeech*, 2004. [Online]. Available: https://api.semanticscholar.org/CorpusID:18880375.

[6] O. Galibert et al., "Named and specific entity detection in varied data: The quæro named entity baseline evaluation," in *International Conference on Language Resources and Evaluation*, 2010. [Online]. Available: https://api.semanticscholar.org/CorpusID:18239707.

[7] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time–frequency weighted noisy speech," *Trans. Audio, Speech and Lang. Proc.*, vol. 19, no. 7, pp. 2125–2136, Sep. 2011, ISSN: 1558-7916. DOI: 10.1109/TASL.2011.2114881. [Online]. Available: https://doi.org/10.1109/TASL.2011.2114881.

[8] A. Graves, *Sequence transduction with recurrent neural networks*, 2012. arXiv: 1211.3711 [cs.NE]. [Online]. Available: https://arxiv.org/abs/1211.3711.

[9] M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, and M. Sonderegger, "Montreal forced aligner: Trainable text-speech alignment using kaldi," in *Interspeech*, 2017. [Online]. Available: https://api.semanticscholar.org/CorpusID:12418404.

[10] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, *Focal loss for dense object detection*, 2018. arXiv: 1708.02002 [cs.CV]. [Online]. Available: https://arxiv.org/abs/1708.02002.

[11] I. Loshchilov and F. Hutter, *Decoupled weight decay regularization*, 2019. arXiv: 1711.05101 [cs.LG]. [Online]. Available: https://arxiv.org/abs/1711.05101.

[12] R. Ardila et al., *Common voice: A massively-multilingual speech corpus*, 2020. arXiv: 1912.06670 [cs.CL]. [Online]. Available: https://arxiv.org/abs/1912.06670.

[13] R. Ardila et al., "Common voice: A massively-multilingual speech corpus," eng, in *Proceedings of the Twelfth Language Resources and Evaluation Conference*, N. Calzolari 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/.

[14] H. Yadav, S. Ghosh, Y. Yu, and R. R. Shah, "End-to-end named entity recognition from english speech," in *Interspeech*, 2020. [Online]. Available: https://api.semanticscholar.org/CorpusID:218862894.

[15] Y. Fu et al., "Aishell-4: An open source dataset for speech enhancement, separation, recognition and speaker diarization in conference scenario," in *Interspeech*, 2021. [Online]. Available: https://api.semanticscholar.org/CorpusID:233181491.

[16] J. E. Hu et al., "Lora: Low-rank adaptation of large language models," *ArXiv*, vol. abs/2106.09685, 2021. [Online]. Available: https://api.semanticscholar.org/CorpusID:235458009.

[17] D. Porjazovski, J. Leinonen, and M. Kurimo, "Attention-based end-to-end named entity recognition from speech," in *Text, Speech, and Dialogue: 24th International Conference, TSD 2021, Olomouc, Czech Republic, September 6–9, 2021, Proceedings*, Olomouc, Czech Republic: Springer-Verlag, 2021, pp. 469–480, ISBN: 978-3-030-83526-2. DOI: 10.1007/978-3-030-83527-9_40. [Online]. Available: https://doi.org/10.1007/978-3-030-83527-9_40.

[18] C. Wang et al., "VoxPopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, C. Zong, F. Xia, W. Li, and R. Navigli, Eds., Online: Association for Computational Linguistics, Aug. 2021, pp. 993–1003. DOI: 10.18653/v1/2021.acl-long.80. [Online]. Available: https://aclanthology.org/2021.acl-long.80/.

[19] A. Bhattacharjee et al., "BanglaBERT: Language model pretraining and benchmarks for low-resource language understanding evaluation in Bangla," in *Findings of the Association for Computational Linguistics: NAACL 2022*, M. Carpuat, M.-C. de Marneffe, and I. V. Meza Ruiz, Eds., Seattle, United States: Association for Computational Linguistics, Jul. 2022, pp. 1318–1327. DOI: 10.18653/v1/2022.findings-naacl.98. [Online]. Available: https://aclanthology.org/2022.findings-naacl.98/.

[20] 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.

[21] J. Wei et al., *Finetuned language models are zero-shot learners*, 2022. arXiv: 2109.01652 [cs.CL]. [Online]. Available: https://arxiv.org/abs/2109.01652.

[22] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. Mcleavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *Proceedings of the 40th International Conference on Machine Learning*, A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, and J. Scarlett, Eds., ser. Proceedings of Machine Learning Research, vol. 202, PMLR, 23–29 Jul 2023, pp. 28 492–28 518. [Online]. Available: https://proceedings.mlr.press/v202/radford23a.html.

[23] G. Ayache, M. Pirchi, A. Navon, A. Shamsian, G. Hetz, and J. Keshet, "Whisperner: Unified open named entity and speech recognition," *ArXiv*, vol. abs/2409.08107, 2024. [Online]. Available: https://api.semanticscholar.org/CorpusID:272599841.

[24] M. Benaicha, D. Thulke, and M. A. T. Turan, *Leveraging cross-lingual transfer learning in spoken named entity recognition systems*, 2024. arXiv: 2307.01310 [cs.CL]. [Online]. Available: https://arxiv.org/abs/2307.01310.

[25] T. Guo et al., *Large language model based multi-agents: A survey of progress and challenges*, 2024. arXiv: 2402.01680 [cs.CL]. [Online]. Available: https://arxiv.org/abs/2402.01680.

[26] L. Wagner, B. Thallinger, and M. Zusag, *Crisperwhisper: Accurate timestamps on verbatim speech transcriptions*, 2024. arXiv: 2408.16589 [cs.LG]. [Online]. Available: https://arxiv.org/abs/2408.16589.

[27] S. Zhou, Z. Li, C. Gong, L. Zhang, Y. Hong, and M. Zhang, "Chinese spoken named entity recognition in real-world scenarios: Dataset and approaches," in *Annual Meeting of the Association for Computational Linguistics*, 2024. [Online]. Available: https://api.semanticscholar.org/CorpusID:271931237.

[28] G. Team et al., *Gemma 3 technical report*, 2025. arXiv: 2503.19786 [cs.CL]. [Online]. Available: https://arxiv.org/abs/2503.19786.

[29] G. Zhu, R. Xiao, H. Wang, Z. Zhu, G. Lyu, and J. Zhao, "Large margin representation learning for robust cross-lingual named entity recognition," in *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, W. Che, J. Nabende, E. Shutova, and M. T. Pilehvar, Eds., Vienna, Austria: Association for Computational Linguistics, Jul. 2025, pp. 4270–4291, ISBN: 979-8-89176-251-0. DOI: 10.18653/v1/2025.acl-long.215. [Online]. Available: https://aclanthology.org/2025.acl-long.215/.