Nwāchā Munā: A Devanagari Speech Corpus and Proximal Transfer Benchmark for Nepal Bhasha ASR

Rishikesh Kumar Sharma, Safal Narshing Shrestha, Jenny Poudel, Rupak Tiwari, Arju Shrestha, Rupak Raj Ghimire, Bal Krishna Bal
Information and Language Processing Research Lab, Kathmandu University
{rishi70612, safalnarsingh, jennypoudel100, rupaktiwari18, arzzustha}@gmail.com,
rughimire@gmail.com, bal@ku.edu.np

Abstract

Nepal Bhasha (Newari), an endangered language of the Kathmandu Valley, remains digitally marginalized due to the severe scarcity of annotated speech resources. In this work, we introduce **Nwāchā Munā** (/nwaːtʃaː munaː/), a newly curated 5.39-hour manually transcribed Devanagari speech corpus for Nepal Bhasha, and establish the first benchmark using script-preserving acoustic modeling. We investigate whether proximal cross-lingual transfer from a geographically and linguistically adjacent language (Nepali) can rival large-scale multilingual pretraining in an ultra-low-resource Automatic Speech Recognition (ASR) setting. Fine-tuning a Nepali Conformer model reduces the Character Error Rate (CER) from a 52.54% zero-shot baseline to 17.59% with data augmentation, effectively matching the performance of the multilingual *Whisper-Small* model despite utilizing significantly fewer parameters. Our findings demonstrate that proximal transfer from Nepali language serves as a computationally efficient alternative to massive multilingual models. We openly release the dataset and benchmarks to digitally enable the Newari community and foster further research in Nepal Bhasha.

Keywords: Automatic Speech Recognition, Nepal Bhasha, Speech Corpus, Low-Resource Speech Recognition

1. Introduction

In the current space of AI, high-resource languages dominate in terms of resource availability while indigenous languages remain digitally marginalized. In a country like Nepal, with 124 official languages, this divide is especially vivid, creating a need for technologies like ASR to bridge the gap. Nepal Bhasha (Newari) serves as a prime example of this imbalance, despite being actively spoken by over 860,000 people in Nepal. Genetically, Nepal Bhasha belongs to the Tibeto-Burman branch of the Sino-Tibetan family (Eberhard et al., 2026). Even with a six-century-old history and a designation as one of the official working languages of Bagmati Province (Post Report, 2024), Nepal Bhasha is currently classified by UNESCO as a "definitely endangered" language (Moseley, 2010). Building a system such as ASR empowers these communities to maintain their linguistic heritage while fully participating in the modern technological landscape.

The modern field of speech recognition has transitioned from fragmented components based on Hidden Markov Models (Rabiner, 1989) to unified end-to-end (E2E) architectures such as Transformer (Vaswani et al., 2017) and Conformer (Gulati et al., 2020). These modern architectures require large datasets to achieve better performance. Recent advancements have yielded robust ASR systems for languages like Nepali and Hindi, but indigenous languages such as Nepal Bhasha remain severely bottlenecked by an extreme scarcity of annotated speech data. This research addresses the fundamental gap by developing a speech corpus for the Newari language—Nwāchā Munā. To further demonstrate the usefulness of this dataset, we establish an initial model benchmark and synthesize modern ASR architectures with cross-lingual transfer learning, leveraging a proximate source language to overcome data constraints.

From this research, we aim to provide a scalable blueprint for other underrepresented languages in the region.

Our contributions are threefold:

1. We release Nwāchā Munā, a carefully curated 5.39-hour Devanagari speech corpus for Nepal Bhasha.
2. We provide the first controlled comparison between proximal transfer (Nepali $\rightarrow$ Newari) and multilingual pretraining (Whisper) in this ultra-low-resource setting.
3. We show that script-preserving proximal transfer can match large multilingual models while requiring substantially fewer parameters and computational resources.

The remainder of this paper is structured as follows: Section 2 reviews related works in low-resource ASR. Section 3 describes the dataset

Proceedings of the Second workshop on Challenges in Processing South Asian Languages (CHiPSAL2026) @ LREC, pages 105–114
16 May 2026.

and methodology including data collection and model training strategies. Section 4 presents the experimental results, discussions and error analysis. Section 5 presents the conclusion followed by ethical statement and limitations.

2. Related works

One major obstacle in building ASR systems for underrepresented languages is the lack of sufficient speech data. Older methods like Hidden Markov Models required less data but struggled to capture intricate acoustic patterns. With time, new architectures like Transformers (Vaswani et al., 2017) and Conformer (Gulati et al., 2020) have outperformed older statistical methods, but they require larger datasets. To address this issue, researchers have increasingly turned to transfer learning. For instance, Cheng et al. (2024) demonstrated clear ASR performance gains for two at-risk Austronesian languages, Amis and Seediq, despite having minimal available speech samples. They achieved this by utilizing a novel data-selection scheme to automatically extract phonetically similar utterances from a broad multilingual corpus to augment their training data. Likewise, significant progress has been made specifically within the South Asian region. The AI4Bharat initiative (Javed et al., 2022) advanced regional ASR by curating over 17,000 hours of raw speech data across 40 Indian languages for continued pre-training.

In the context of Nepali ASR, early efforts relied on Hidden Markov Models (HMM) (Ssarma et al., 2017) for isolated word recognition, which struggled with continuous speech. The transition to deep learning began with Regmi et al. (2019), they implemented Recurrent Neural Networks (RNN) with Connectionist Temporal Classification (CTC) loss. This was further refined by Bhatta et al. (2020) with a hybrid CNN-GRU-CTC model. Furthermore, Regmi and Bal (2021) implemented end-to-end speech recognition approach for the Nepali language. To address training instability in deeper networks, Dhakal et al. (2022) proposed a CNN-BiLSTM-ResNet architecture, achieving a CER of 17.06%. Concurrently, research began shifting towards data efficiency and cross-lingual techniques. For example, Bansal et al. (2020) explored an acoustic-phonetic approach, demonstrating that leveraging monolingual and cross-lingual information could improve performance in low-resource settings. Joshi and Shrestha (2023) applied Self-Attention Networks (SAN) to capture long-range dependencies better than recurrent models. Building on these modern architectures, Paudel et al. (2023) successfully applied a combination of CNNs and Transformers for large vocabulary continuous speech recognition.

Current state-of-the-art systems have fully embraced E2E Transformer-based architectures. Poudel et al. (2025) introduced `NepConformer`, a Conformer-based Nepali ASR model achieving a CER of `6.01%`. Recent focus has also turned to data selection, tokenization and resource efficient strategies. Ghimire et al. (2023a) proposed an Active Learning framework to select the most informative samples for speech annotation. Another work (Ghimire et al., 2023b), demonstrated that using linguistically motivated sub-word units (syllables) alongside active sampling significantly outperforms standard methods. Work done by Ghimire et al. (2025) showed that using adapter-based Parameter-Efficient-Fine-Tuning (PEFT) alone improves Word Error Rate (WER) by 19% compared to `Whisper-large-v2`. Extending the utility of efficient adaptation, Pantha et al. (2025) demonstrated the effectiveness of PEFT for speech personalization among Nepali speakers. Most recently, Ghimire et al. (2023b) implemented new complementary loss function for addressing grammatical consistency during model training and fine-tuning. Their Rule-Based Character Constituency Loss (RBCCL) help to reduce Word Error Rates (WER) from 47.1% to 23.41% by penalizing grammatically impossible character sequences during training.

While Nepali has seen consistent growth in the field of ASR, research on Nepal Bhasha remains sparse. A notable exception is the work by Meelen et al. (2024), who presented the first Newari ASR, where they achieved a CER of 12%. Their evaluation is conducted in Romanized transliteration, this fails to preserve Devanagari orthography. Their study highlighted that while transfer learning is effective, the lack of standardized orthography and limited data results in higher WER compared to Nepali. They emphasized the need for "orthography standardization" and data augmentation techniques like SpecAugment (Park et al., 2019) to make ASR viable for these endangered languages.

Taken together, these findings highlight a persistent gap: prior work lacks ASR systems trained on the Devanagari script using carefully curated low-resource datasets. This research clears these gaps by developing Nwāchā Munā. We utilise Devanagari script for transcribing speech of Nepal Bhasha. We also use data augmentation techniques to train robust ASR models using various cross-lingual transfer strategies and report CER for each technique.

3. Methodology

3.1. Data Collection Technique

The development of the Nepal Bhasha ASR required a carefully designed data collection process to ensure authenticity and reliability. Both textual and acoustic resources were gathered considering language and speaker diversity. The objective was to construct a dataset that reflects natural language usage while maintaining consistency and structural integrity. The following subsections describe the procedures adopted for textual and audio data acquisition.

3.1.1. Textual Data Acquisition

To establish a baseline for the Nepal Bhasha ASR system, a comprehensive textual corpus was compiled from different sources of digital and print media. The data collection strategy included the use of both formal and everyday language to capture the full range of vocabulary. Digital data were extracted from Nepal Bhasha Wikipedia [1] and the OSCAR (Open Super-large Crawled Aggregated coRpus) (Ortiz Suárez et al., 2020) dataset while print-based resources included regional newspapers, literary manuscripts from the Bloom Library [2], and primary school textbooks. In addition to the formal texts obtained, a targeted set of daily conversational sentences was curated manually to incorporate texts that were previously missing. The raw text was rigorously pre-processed where non-target language tokens, numerical digits, and idiosyncratic symbols were manually removed to ensure high fidelity and relevance. Table 1 shows an example of text normalization, where raw text containing Devanagari numerals and symbols is converted into a phonetic, standardized form for acoustic modeling.

Table 1: Example of Text Normalization for Nepal Bhasha ASR

Condition	Sentence Structure
Before	''छगु ब्वसय् २१% ब्व दु।'' chagu bwasay 21% bwa du *The sentence contains 21 as a number.*
After	छगु ब्वसय् नीछगू सयेक ब्व दु chagu bwasay nīchagū sayeka bwa du *The sentence contains 21 in textual form*

The overall text corpus statistics and sentence length distribution are presented in Table 2 and 3:

Table 2: Nwāchā Munā Corpus Overview

Metric	Count
Total Sentences	5,727
Total Words	27,644
Unique Words in Corpus	8,599

Table 3: Sentence Length Distribution in Nwāchā Munā Corpus

Sentence Length	Count
Short (0–5 words)	4,323
Medium (6–10 words)	1,073
Long (> 10 words)	331

3.1.2. Audio Data Acquisition

Primary acoustic data was gathered through a dual-modal strategy involving original field recordings and the transcription of existing web-based audio, ensuring the model's exposure to varied acoustic environments. To maintain technical consistency across the dataset, all audio was standardized to a 16 KHz sampling rate and stored in a mono-channel WAV format, providing the necessary fidelity for deep learning-based feature extraction. For the recording phase, 18 native speakers from Banepa, Dhulikhel, Panauti, and Patan volunteered based on criteria such as native fluency and linguistic knowledge. While Newari exhibits location-based dialectal variations, the use of a standardized text across all recordings ensured that the resulting differences remained primarily tonal. To enhance the model's robustness against speaker-dependent variability, the data was stratified by age and gender. Audio recordings were captured using built-in smartphone microphones in an open environment with minimal acoustic disturbance. This field collection of 4 hours and 21 minutes of speech was further supplemented by approximately one hour of web-sourced audio (Meelen et al., 2024), which was originally in Romanized form and subsequently transliterated into Devanagari script by community members.

The overall characteristics of the Nwāchā Munā speech corpus are summarized in Tables 4-6. Table 4 presents the corpus-level statistics, including the total number of utterances, total duration, and mean utterance length. Table 5 details the age distribution of the same participant set, while Table 6 provides the gender distribution of the 18 speakers contributing to the corpus.

[1] https://new.wikipedia.org/wiki/
[2] https://bloomlibrary.org/language:new

Table 4: Overview of the Nwāchā Munā

Metric	Value
Total Utterances	5,727
Total Duration	5.39 hours
Mean Utterance Length	3.39 sec

Table 5: Age Distribution of Speakers in the Nwāchā Munā Corpus

Age Group	Number of Speakers
16–25	6
26–30	2
30+	10
Total Speakers	18

Table 6: Gender Distribution of Speakers in the Nwāchā Munā Corpus

Gender	Number of Speakers
Male	10
Female	8
Total Speakers	18

3.2. Model Training Strategies

We adopted a comparative experimental framework to systematically evaluate cross-lingual transfer strategies for Newari automatic speech recognition. Our central hypothesis is that acoustic and orthographic proximity between Nepali and Newari enables effective encoder reuse, thereby reducing the need for large multilingual pretraining. To test this hypothesis, we structured our investigation into three core components: 1) acoustic modeling, 2) language modeling and decoding, and 3) semi-supervised learning.

In the domain of acoustic modeling, we evaluated distinct training strategies starting with a zero-shot evaluation of the pre-trained `NepConformer` (Poudel et al., 2025), followed by supervised fine-tuning of NepConformer, and fine-tuning of a massive multilingual model, `Whisper-Small` (244M parameters) on the collected Newari dataset. To test the adaptability of pre-trained acoustic representations, we further distinguished between standard full model fine-tuning and a decoder-only fine-tuning strategy. We also investigated the impact of data augmentation techniques in improving the performance and robustness of the models.

Finally, to address low-resource constraints in decoding, we explored language modeling strategies by incorporating an external KenLM-based n-gram model via shallow fusion to enhance transcription accuracy.

3.2.1. Acoustic Modeling Strategies

The zero-shot evaluation of NepConformer served as a baseline to assess cross-lingual generalization without any exposure to Newari during training. This setup enabled us to measure the extent to which acoustic and linguistic representations learned from Nepali transfer to a closely related but distinct language. Subsequently, we employed transfer learning through supervised fine-tuning on the Newari corpus to adapt the NepConformer model to the target language. This allowed us to measure the gains obtained from domain-specific adaptation compared to zero-shot performance. To test the adaptability of pre-trained acoustic representations, we further distinguished between standard full model fine-tuning and a decoder-only fine-tuning strategy, where the encoder parameters were frozen to assess the sufficiency of source-language acoustic features. Finally, we fine-tuned Whisper-Small to evaluate whether broad cross-lingual pre-training yielded stronger transfer capabilities than a monolingual Nepali trained model.

To combat data scarcity in low-resource Nepal Bhasha ASR, a dual-stage data augmentation strategy was employed. By combining static (offline) and dynamic (online) methods, we introduced diverse acoustic variations to improve model robustness and prevent overfitting. This approach acts as a regularizer, forcing the model to learn invariant phonetic features rather than memorizing the specific acoustic signatures of the limited training set.

3.2.2. Language Modeling and Decoding

To enhance linguistic modeling under low-resource constraints, we used two approaches: decoder-only fine-tuning and shallow fusion with an external language model (KenLM) (Heafield, 2011).

In the decoder-only setup, the encoder parameters were frozen, and only the prediction network and joint network were updated during fine-tuning. By applying the phonetic similarities between Nepali and Newari, this strategy conserves the transferable acoustic representations obtained during pre-training while allowing the decoder to adapt specifically to Newari linguistic patterns.

Independent of decoder adaptation, we also integrated an external n-gram language model trained using KenLM and applied shallow fusion

during inference. This introduces explicit linguistic priors to capture word-level dependencies that the limited paired data might miss. During beam search decoding, the log-probabilities from the acoustic model and the external language model were linearly combined (Toshniwal et al., 2018), with the fusion weight tuned on the development set.

For the baseline system, beam search decoding approximates the most likely sequence:

$$y^* = \arg\max_y \log p(y \mid x)$$

Under shallow fusion, we incorporate the external language model via log-linear interpolation with a word insertion bonus:

$$y^* = \arg\max_y \{\log p(y \mid x) + \alpha \log p_{LM}(y) + \beta |y|_w\}$$

where α controls the language model weight, β is the word insertion bonus, and $|y|_w$ denotes the number of words in the hypothesis y.

Decoder-only fine-tuning improves the model's language understanding by updating only the higher-level components while keeping the encoder fixed, which also reduces computational cost. This setup allows us to observe the effect of adapting the decoder alone in a low-resource setting. While, shallow fusion adds an external language model during decoding without changing the acoustic model itself. By evaluating these two methods independently, we can clearly measure the contributions of internal decoder adaptation and external statistical language modeling to CER, WER, and improve transcription consistency in low-resource settings.

3.2.3. Semi-Supervised Learning

Due to the availability of limited data for Nepal Bhasha, we adopted a semi-supervised pseudo-labeling framework. An unlabeled corpus was collected from publicly available broadcast sources, including radio, podcasts, and television, and segmented into shorter utterances. To mitigate memory issues and alignment degradation in CTC-based training, excessively long utterances were discarded. The remaining audio was transcribed using our best intermediate acoustic model to generate pseudo-labels.

To ensure quality, we applied rigorous filtering. Single-word transcripts were removed, as they were typically noise-induced fragments rather than valid words. We also computed average prediction confidence scores and applied a threshold filter to reduce error reinforcement before incorporating the pseudo-labeled data into our training pipeline.

4. Results and Discussion

4.1. Experimental Setup

The dataset was divided into 80% for training, 10% for validation, and 10% for testing. Due to the limited size of the dataset across 18 participants, speaker representation was maintained in all splits to prioritize overall acoustic diversity rather than enforcing strict speaker independence. Table 7 summarizes the Newari speech corpus split across training, validation, and test sets, showing the number of hours and utterances for each partition.

Table 7: Dataset split for Newari speech corpus

Split	Hours	Utterances
Train	4.31	4575
Validation	0.54	576
Test	0.54	576

For the Conformer model, static augmentation expanded the training set by a factor of 5 via speed perturbation ($0.9\times, 1.1\times$), volume randomization, and noise injection, resulting in 23.05 total hours of audio. Dynamic augmentation was applied with a probability of $p = 0.5$, involving time stretching (rate $0.8\times$–$1.25\times$), pitch shifting (± 4 semitones), and Gaussian noise injection (amplitude 0.001–0.015).

The fine-tuned Nepal Bhasha ASR model utilizes a `Conformer-CTC` architecture with a Byte Pair Encoding (BPE) tokenizer of vocabulary size 128. The 18-layer Conformer encoder has a d-model dimension of 256, 4 attention heads, and a convolution kernel size of 31. Regularization includes a dropout rate of 0.1 across the encoder and attention modules, alongside SpecAugment with 2 frequency masks (width 27) and 2 time masks (width 70). Optimization is performed using AdamW with a learning rate of 1×10^{-4} and a Cosine Annealing schedule with 3,000 warmup steps. T4 GPU with 16GB vRAM and 29GB of system RAM was used for training in Kaggle environment with total training time of 3 hours for standard data and 12 hours for augmented data.

The `Whisper-Small` model was fine-tuned for 10 epochs using a sequence-to-sequence objective, with the decoder explicitly forced to the Nepali language token in order to leverage script compatibility. Optimization was performed using AdamW with a cosine learning rate decay (peak 1×10^{-5}, 10% warmup) and an effective batch size of 16, achieved via gradient accumulation under FP16 mixed-precision. To further ensure stability, a dropout rate of 0.2 to all attention and hidden layers were applied. A T4-GPU with 15GB vRAM was

used for training this model in Google Colab environment with total training time of 4.5 hours.

For language modeling and decoding, we performed a grid search on the development set to identify optimal decoding hyperparameters across all models. We explored different configurations of language model weight (α), word insertion bonus (β), and beam width. Based on this systematic evaluation, a 5-gram KenLM with $\alpha = 0.4$, $\beta = 1.5$, and a beam width of 128 was selected, as it consistently yielded the best performance improvements. The language model was trained on a text corpus consisting of approximately 51k utterances and 478k word tokens, sourced from Nepal Bhasha Wikipedia. Notably, this dataset was reserved solely for language modeling and is distinct from the audio transcription data.

For the semi-supervised learning experiments, an initial unlabeled corpus of 13.65 hours was collected. During the filtering phase, utterances longer than 15 seconds were discarded. After transcribing the audio with our intermediate acoustic model, a 70% confidence threshold was applied to the predictions to ensure pseudo-label quality. Following this rigorous filtering process, 9.33 hours of high-quality pseudo-labeled data remained for experimental training.

4.2. Findings and Analysis

The performance of the models was evaluated using CER as the primary metric. The experiments were conducted on the Nwāchā Munā dataset, while additional evaluations incorporated augmented data and unlabelled speech to examine the effect of data expansion and semi-supervised training. The quantitative results are summarized in Table 8, where only the best-performing models at their optimal training epochs are reported.

Due to differences in script representation, normalization, evaluation protocol and the unavailability of their exact test split, a direct numerical comparison with (Meelen et al., 2024) is not feasible. Instead, our experimental results demonstrate that transferring representations from a neighboring language with a similar script is a viable alternative to massive multilingual pre-training for low-resource ASR, suggesting that linguistic proximity can potentially outweigh model scale in ultra-low-resource South Asian scenarios.

Table 9 shows the orthographic and phonological comparison of the Nepali and Newari Language. Table 10 presents a comparison of model size and computational resource requirements, highlighting that `NepConformer` is significantly more lightweight, with only 30.54 million parameters and lower VRAM usage (8.56 GB) compared to `Whisper-Small`, which requires substantially higher usage (12.38 GB).

Despite possessing significantly fewer parameters and lacking the huge data backing available in multilingual models like Whisper, the `NepConformer` model achieved a baseline CER of 18.72%, effectively matching the performance of the fine-tuned `Whisper-Small` model (18.76%). This parity suggests that the acoustic and orthographic overlap between Nepali and Newari facilitates robust feature extraction, minimizing the need for the extensive capacity of large-scale models. Furthermore, the application of data augmentation proved critical in this data-scarce regime, yielding the state-of-the-art performance of 17.59% CER for `NepConformer` and 17.88% for `Whisper-Small`, highlighting that augmenting scarce audio data can effectively overcome dataset limitations and drive substantial performance gains.

Conversely, a closer examination of our training configurations reveals critical insights into the

Table 8: Summary of Results for Evaluated Modeling Strategies

Strategies (model details)	**Dataset**	**CER (%)**	**WER (%)**
Zero-Shot NepConformer	-	52.54	98.68
Semi-supervised NepConformer	5.39 hours labelled + 9.33 hours Pseudo labeled	19.83	58.28
NepConformer + 5-gram KenLM (Shallow fusion, $\lambda = 0.4$, $\beta = 1.5$, beam 128)	5.39 hours labelled	19.75	**48.46**
Decoder-only NepConformer (Encoder frozen)	5.39 hours labelled	18.77	53.47
Whisper-Small (Fine-tuned on base data)	5.39 hours labelled	18.76	51.44
NepConformer (Fine-tuned on base data)	5.39 hours labelled	18.72	54.34
Whisper-Small + Augmentation (Time stretch, pitch shift, Gaussian noise)	5.39 hours labelled	17.88	48.80
NepConformer + Augmented Data (speed perturbation ($0.9\times, 1.1\times$), volume randomization, and noise injection)	23.05 hours labelled	**17.59**	52.61

Table 9: Orthographic and Phonological Comparison of Nepali and Newari

Features	Nepali	Newari
Script	Devanagari	Devanagari
Nasalization Markers	Yes	Yes (more enhanced)
Agglutination	Moderate	High
Halant Usage	Standard	High Morphological Load

Table 10: Computational Resource and Parameter Comparison

Model	Parameter (Million)	VRAM Used (GB)
NepConformer	30.54	8.56
Whisper(small)	244	12.38

adaptation process. The poor zero-shot performance (52.54% CER) confirms that despite script sharing, the phonological distinctiveness of Newari requires explicit fine-tuning. The *decoder-only* fine-tuning of `NepConformer` yielded results (18.77% CER) nearly identical to full model fine-tuning, indicating that the pre-trained Nepali encoder features are sufficiently generalized to encode Newari speech without modification.

Additionally, the integration of the KenLM n-gram language model was evaluated across all models during decoding. The best results were observed with the NepConformer fine-tuned model, where WER was reduced by approximately 11.7% relative to the baseline. At the same time, the CER increased slightly by 1.37%, as the language model tends to favor standard spellings, which can override phonetically correct but unusual or local forms in Newari speech. This happens partly because the text used to train the language model doesn't always match how people actually speak. Overall, this shows the trade-off between making words more consistent and keeping the natural variations of this low-resource, agglutinative language.

Finally, in the semi-supervised learning approach, incorporating the pseudo-labeled data with the original labeled training set degraded performance, increasing the CER to 19.83% compared to the 17.59% baseline. A qualitative analysis of a random sample of these pseudo-labels revealed three primary sources of error: extensive code-mixing (alternating with Nepali or English), frequent overlapping conversational speech, and highly variable acoustic environments (e.g., background music, distant microphones). Ultimately, this decline highlights the critical challenge that the severe domain shift introduced by uncurated data can completely overshadow the benefits of increased training volume. This suggests that strict domain alignment, speaker diarization, and rigorous filtering are significantly more crucial than raw data quantity for endangered-language ASR.

4.3. Error Analysis

Newari's rich morphological structure, characterized by productive affixation and concatenated morphemes, makes character-level analysis particularly informative. Although the model frequently predicts individual characters correctly, it struggles to assemble them into accurate representations. As a result, the CER remains relatively low, while the WER is substantially higher, reflecting the difficulty of modeling agglutinative formations and complex orthographic concatenations. This is also observed in other work (D K et al., 2025).

Qualitative analysis indicates that transcription errors are primarily associated with specific diacritics: हलन्त (◌्) (vowel suppressor), अनुस्वार (◌ं) (nasalization), चंद्रबिंदु (◌ँ) (nasal vowel), and विसर्ग (◌ः) (aspiration). These markers encode essential phonetic distinctions, including nasalization, short stops, and breathy releases; their omission alters the acoustic-orthographic correspondence. As further illustrated in Table 11, these findings underscore the importance of language-specific grapheme modeling and the development of corpora encompassing both formal and conversational usage.

Table 11: Error Frequency Table

Error Type	% of total errors
Word Boundary Errors	35%
Halant/Cluster Errors	25%
Nasalization & Anusvara Confusion	20%
Lexical Substitution	15%
Function Word & Particle Error	5%

The sample qualitative error are presented in Table 12. This distribution shows that the primary challenges arise not from isolated character misrecognition but from morphological segmentation complexity, diacritic modeling, and the compounding effect of Devanagari's script characteristics.

Table 12: Qualitative Error Analysis of Nwāchā Munā (underline in prediction column represents the error segment)

No.	Reference	Prediction	Error Analysis
(i)	तस्सकं न्ह्याइपुसेच्वं सीम बत्तिं चैं छन्दत म्याहला च्वंगु tassakaṃ nhyāipusecvaṃ sīma battiṃ caiṃ chandata myāhalā cvaṃgu *It was a lot of fun sitting even in the dim light, singing your song*	तस्सकां न्येयेपुसं सिमापत्ति चैं छंत मां ला च्वं tassakāṃ nyeyepusaṃ simāpatta caiṃ chaṃta māṃ lā cvaṃ	Phonetic substitutions and word distortions; nasal markers and vowels altered within agglutinated structures.
(ii)	दइ धका नं सिल हे daï dhakā naṃ sila he *They say it is known that it will fall*	दइ धका: नं सिल हे daï dhakāḥ naṃ sila he	Insertion of (ः) (Visarga), reflecting minor phonetic misalignment.
(iii)	फोहोर जुलकि लाकां सिले माः phohora julaki lākāṃ sile māḥ *If the shoe is dirty, it must be washed*	पर जुलकि लाका सिने माः para julaki lākā sine māḥ	Nasal suffix deletion and lexical substitution; final (ं) dropped, with minor token distortion.
(iv)	आः थ्व सु āḥ thva su *Now, who is this*	आाख्वसु āākhvasu	Word boundary deletion; independent tokens merged due to agglutinative structure.

5. Conclusion

In this work, we developed resources to help bridge the digital divide for Nepal Bhasha (Newari) by curating a high-quality 5.39-hour transcribed speech corpus, with all transcripts encoded in the Devanagari script. Using this resource, we systematically evaluated cross-lingual transfer learning strategies, benchmarking `NepConformer` along with a multilingual `Whisper-Small` baseline. Our experiments demonstrate substantial gains from supervised fine-tuning and data augmentation, reducing the CER from a high zero-shot baseline to below 18%. The frozen-encoder fine-tuning strategy further enabled stable and effective adaptation in this ultra-low-resource setting. Moreover, shallow fusion with a KenLM n-gram language model enhances lexical regularity and contributes to a reduction in WER, even when character-level improvements remain marginal, indicating improved word-level consistency in decoded outputs. Beyond Nepal Bhasha, our findings indicate that intra-regional transfer learning within South Asian language clusters may offer a computationally efficient pathway for scaling ASR to other endangered and minority languages

Data and Code Availability The training scripts and source code are available on GitHub (https://github.com/ilprl/nwacha-muna). The model weights and audio datasets are hosted on Hugging Face (https://huggingface.co/collections/ilprl-docse/nwacha-muna).

6. Ethics Statement and Limitations

This work offers meaningful societal and technological benefits for the Nepal Bhasha speaking community. By providing an open-source Devanagari speech corpus and establishing strong ASR benchmarks, the resources developed pave the way for voice-driven AI products that can significantly enhance digital accessibility. Furthermore, this research plays an important role in building extensive speech archives, allowing endangered language communities to actively document and preserve their linguistic heritage.

The dataset was collected through voluntary participation from the community, involving native speakers across diverse age groups and genders to ensure representative coverage. For the recorded audio, consent was obtained from all volunteers, and no sensitive or personally identifiable information (PII) was intentionally included in the spoken transcripts. Nonetheless, we acknowledge that models trained on limited data may reflect embedded societal biases or result in acoustic-phonetic misrepresentations, particularly in ultra-low-resource settings. Because of these limitations, the baseline ASR system developed in this study should not be deployed in critical applications without rigorous human verification. The primary objective of this work remains the promotion of linguistic inclusion and digital support for under-resourced language communities.

Despite the promising benefits, this work faces several constraints related to ultra-low-resource settings. Primarily, the curated Nwāchā Munā

dataset, while high-quality, remains relatively small compared to standard ASR benchmarks. Our current evaluation focuses predominantly on read speech and formal sentences, which may not fully capture the acoustic variability, disfluencies, and rapid speaking rates encountered in everyday spontaneous conversations. Also, external n-gram language model utilized during shallow fusion is constrained by domain mismatches and possesses limited coverage of natural, conversational linguistic patterns.

7. Bibliographical References

Shweta Bansal, Shweta Sinha, and Shyam S. Agrawal. 2020. Acoustic-phonetic approach for ASR of less resourced languages using monolingual and cross lingual information. In *Proceedings of the 1st Joint SLTU and CCURL Workshop (SLTU-CCURL 2020)*, pages 167–171. European Language Resources Association (ELRA).

Bharat Bhatta, Basanta Joshi, and Ram Krishna Maharjhan. 2020. Nepali speech recognition using CNN, GRU and CTC. In *Proceedings of the 32nd Conference on Computational Linguistics and Speech Processing (ROCLING 2020)*, pages 238–246. The Association for Computational Linguistics and Chinese Language Processing.

Yao-Fei Cheng, Li-Wei Chen, Hung-Shin Lee, and Hsin-Min Wang. 2024. Exploring the impact of data quantity on asr in extremely low-resource languages.

Thennal D K, Jesin James, Deepa Padmini Gopinath, and Muhammed Ashraf K. 2025. Advocating character error rate for multilingual ASR evaluation. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 4941–4950, Albuquerque, New Mexico. Association for Computational Linguistics.

Manish Dhakal, Arman Chhetri, Aman Kumar Gupta, Prabin Lamichhane, Suraj Pandey, and Subarna Shakya. 2022. Automatic speech recognition for the nepali language using CNN, bidirectional LSTM and ResNet. In *2022 International Conference on Inventive Computation Technologies (ICICT)*, pages 515–521. IEEE.

David M. Eberhard, Gary F. Simons, and Charles D. Fennig. 2026. *Ethnologue: Languages of the World*, twenty-ninth edition. SIL International, Dallas, Texas. Online version: http://www.ethnologue.com.

Rupak Raj Ghimire, Bal Krishna Bal, and Prakash Poudyal. 2023a. Active learning approach for fine-tuning pre-trained ASR model for a low-resourced language: A case study of Nepali. In *Proceedings of the 20th International Conference on Natural Language Processing (ICON)*, pages 82–89, Goa University, Goa, India. NLP Association of India (NLPAI).

Rupak Raj Ghimire, Bal Krishna Bal, Balaram Prasain, and Prakash Poudyal. 2023b. Pronunciation-aware syllable tokenizer for Nepali automatic speech recognition system. In *Proceedings of the 20th International Conference on Natural Language Processing (ICON)*, pages 36–43, Goa University, Goa, India. NLP Association of India (NLPAI).

Rupak Raj Ghimire, Prakash Poudyal, and Bal Krishna Bal. 2025. Improving accuracy of low-resource asr using rule-based character constituency loss (rbccl). In *Proceedings of the First Workshop on Challenges in Processing South Asian Languages (CHiPSAL)*, pages 61–70. Association for Computational Linguistics.

Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, and Ruoming Pang. 2020. Conformer: Convolution-augmented transformer for speech recognition. In *Proceedings of Interspeech 2020*, pages 5036–5040. ISCA.

Kenneth Heafield. 2011. KenLM: Faster and smaller language model queries. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 187–197, Edinburgh, Scotland. Association for Computational Linguistics.

Tahir Javed, Sumanth Doddapaneni, Abhigyan Raman, Kaushal Santosh Bhogale, Gowtham Ramesh, Anoop Kunchukuttan, Pratyush Kumar, and Mitesh M. Khapra. 2022. Towards building ASR systems for the next billion users. In *Proceedings of the Thirty-Sixth AAAI Conference on Artificial Intelligence (AAAI-22)*.

Basanta Joshi and Rupesh Shrestha. 2023. Nepali speech recognition using self-attention networks. *International Journal of Innovative Computing, Information and Control*, 19(6):1769–1784.

Marieke Meelen, Alexander O'Neill, and Rolando Coto-Solano. 2024. End-to-end speech recognition for endangered languages of nepal. In *Proceedings of the 7th Workshop on Computational Methods for Endangered Languages (ComputEL-7)*, pages 83–93. Association for Computational Linguistics.

Christopher Moseley, editor. 2010. *Atlas of the World's Languages in Danger*, 3rd edition. UNESCO Publishing, Paris.

Pedro Javier Ortiz Suárez, Laurent Romary, and Benoît Sagot. 2020. A monolingual approach to contextualized word embeddings for mid-resource languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1703–1714, Online. Association for Computational Linguistics.

Kiran Pantha, Rupak Raj Ghimire, and Bal Krishna Bal. 2025. Speech personalization using parameter efficient fine-tuning for Nepali speakers. In *Proceedings of the 5th Conference on Language, Data and Knowledge: Fifth Workshop on Language Technology for Equality, Diversity, Inclusion*, pages 190–199, Naples, Italy. Unior Press.

Daniel S. Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D. Cubuk, and Quoc V. Le. 2019. Specaugment: A simple data augmentation method for automatic speech recognition. In *Proceedings of Interspeech 2019*, pages 2613–2617. ISCA.

Shishir Paudel, Bal Krishna Bal, and Dhiraj Shrestha. 2023. Large vocabulary continous speech recognition for Nepali language using CNN and transformer. In *Proceedings of the 4th Conference on Language, Data and Knowledge*, pages 328–333, Vienna, Austria. NOVA CLUNL, Portugal.

Post Report. 2024. Nepali, tamang, nepal bhasha official languages of bagmati. *The Kathmandu Post*. Accessed: 2026-02-26.

Jenny Poudel, Ankit Dahal, Rishikesh Kumar Sharma, Rupak Tiwari, Rupak Raj Ghimire, and Bal Krishna Bal. 2025. Nepconformer: A conformer-based nepali automatic speech recognition system. In *International Conference on Computing and Machine Learning*, pages 167–178. Springer.

Lawrence R. Rabiner. 1989. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286.

P Regmi, A Dahal, and B Joshi. 2019. Nepali speech recognition using rnn-ctc model. *International Journal of Computer Applications*.

Sunil Regmi and Bal Krishna Bal. 2021. An end-to-end speech recognition for the Nepali language. In *Proceedings of the 18th International Conference on Natural Language Processing (ICON)*, pages 180–185, National Institute of Technology Silchar, Silchar, India. NLP Association of India (NLPAI).

Manish K. Ssarma, Avaas Gajurel, Anup Pokhrel, and Basanta Joshi. 2017. Hmm based isolated word nepali speech recognition. In *2017 International Conference on Machine Learning and Cybernetics (ICMLC)*, volume 1, pages 1–6. IEEE.

Shubham Toshniwal, Anjuli Kannan, Chung-Cheng Chiu, Yonghui Wu, Tara N Sainath, and Karen Livescu. 2018. A comparison of techniques for language model integration in encoder-decoder speech recognition. In *2018 IEEE Spoken Language Technology Workshop (SLT)*, pages 369–375. IEEE.

Ashish Vaswani, Llion Jones, Noam Shazeer, Niki Parmar, Aidan N. Gomez, Jakob Uszkoreit, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30.