International Journal of Computer Applications (0975 - 8887)
Volume 187 - No.152, November 2026

ShrutiParichhya : A Nepali ASR Evaluation Dataset and Error Analysis of Fine-Tuned ASR Models

Rupak Raj Ghimire
ILPRL, KU[1]

Rupak Tiwari
ILPRL, KU

Rishikesh Kumar Sharma
ILPRL, KU

Jenny Poudel
ILPRL, KU

Safal Narshing Shrestha
ILPRL, KU

Prakash Poudyal
ILPRL, KU

Bal Krishna Bal
ILPRL, KU

ABSTRACT

Automatic Speech Recognition (ASR) for Nepali has improved through fine-tuning of large multilingual speech models. Most studies evaluate these models on a clean and read speech corpus such as OSLR54, and report only Character Error Rate (CER) and Word Error Rate (WER). These metrics show how often a model fails, but not where and why. It also remains unclear how the models perform on naturally recorded speech. This paper addresses both gaps. Whisper-Small and Wav2Vec2-BERT are fully fine-tuned, and MMS-1b-all is fine-tuned with adapters, on OSLR54. Wav2Vec2-BERT performs best on the OSLR54 validation split (WER 7.91%, CER 1.48%). All three models are then evaluated on *ShrutiParichhya* , a new 5-hour corpus of naturally recorded Nepali with 4,428 utterances, 36 speakers, and per-utterance SNR labels. Wav2Vec2-BERT again scores best, but its WER rises to 52.69% and its CER to 20.66%. Error rates change little from the Clean to Moderate SNR tiers, which hold 99.2% of the utterances. The errors concentrate in merged word boundaries, deleted conjunct halants, vowel-length and sibilant confusions, short function words, and digit output. MMS-1b-all merges words most often, and Whisper-Small produces hallucination loops. These errors are consistent with a register mismatch between read training speech and spontaneous test speech. *ShrutiParichhya* differs from OSLR54 in several factors at once, so register was not isolated from model capacity.

Keywords

Nepali ASR, Low-resource Speech Recognition, Signal-to-Noise-Ratio(SNR), Devanagari Script, Error Analysis, Out-of-distribution Evaluation

1. INTRODUCTION

Nepali is an Indo-Aryan language written in the Devanagari script. It is the official language of Nepal and is spoken by 13.08 million people as a first language [21]. Its spelling follows Sanskrit etymology rather than modern pronunciation. As a result, several letters that are written differently sound the same or nearly the same in everyday speech. Examples are the sibilants , , and , and the short and long vowel signs and . An ASR system cannot always choose the correct spelling from the audio alone.

Several recent publications ([10, 14, 18]) show that Nepali ASR research has evolved from early isolated-word systems based on Hidden Markov Models (HMMs) to modern end-to-end (E2E) deep learning based architectures. Current state-of-the-art Nepali ASR models mostly build on large self-supervised multilingual pretrained models (e.g., wav2vec 2.0, Whisper, MMS) and Conformer architectures.

Recent work on Nepali ASR fine-tunes large pretrained multilingual models, such as Whisper and MMS [20, 19]. Parameter-efficient methods such as LoRA and DoRA lower the cost of adapting these models [14]. Despite the published work, three challenges remain. First, public Nepali speech corpora consist mostly of read speech recorded under controlled conditions. Second, Nepali spelling keeps Sanskrit-derived distinctions that speakers no longer pronounce, so the written form is hard to predict from audio alone. Third, real recordings contain noise and accent variation.

Recent studies report results on the OSLR54, FLEURS, and Common Voice speech corpora [23, 17]. All three contain read speech, so it is unclear how fine-tuned models perform on naturally recorded, cross-domain speech. Most studies also report only CER and WER, which show how often a model fails but not why.

This paper addresses both gaps with the following contributions:

(1) Three commonly used multilingual models are fine-tuned on OSLR54.

(2) A new 5-hour evaluation corpus of naturally recorded Nepali with per-utterance SNR annotation, *ShrutiParichhya* , is prepared, and the fine-tuned models are evaluated on it.

(3) The transcription errors are analysed by acoustic condition and by linguistic category.

[1]Information and Language Processing Research Lab, Department of Computer Science and Engineering, Kathmandu University, Dhulikhel, Nepal

Reliable evaluation on real speech data is a necessary step before Nepali ASR can be deployed outside controlled settings.

2. RELATED WORKS

2.1 Nepali Speech Resources

Structured Nepali language resources began with the Nepali National Corpus [27]. Two OpenSLR corpora are now widely used. OSLR54 contains about 157,000 crowd-sourced utterances and supports ASR training [13]. OSLR43 is a studio-recorded corpus for TTS [25]. Multilingual resources add further Nepali speech. Common Voice collects speech read by volunteers [1]. FLEURS provides read speech for evaluation [8].
The corpora in Table 1 contain read or prompted speech. IndicVoices adds extempore and conversational speech for the 22 scheduled Indian languages, which include Nepali [12]. To the authors' knowledge, no public corpus pairs natural Nepali speech with per-utterance SNR labels. *ShrutiParichhya* addresses this gap.

2.2 Classical and Neural Nepali ASR

Early Nepali ASR relied on Hidden Markov Models (HMMs) for isolated and continuous speech [26, 4]. Subsequent end-to-end neural approaches eliminated pronunciation lexicons, progressing through CTC-based RNNs, CNN hybrids, and attention architectures [15, 5, 22, 16]. Recent efforts focus on output-unit optimizations like syllable-aware tokenization [10] and modern encoder baselines; notably, NepConformer [18] achieved 6.01% CER and 23.96% WER on OSLR54.

2.3 Pretrained Multilingual Speech Models

Large pretrained models now dominate low-resource ASR. Wav2Vec 2.0 learns speech representations from unlabeled audio [2]. XLS-R extends this approach to many languages [7]. MMS scales it to more than 1,100 languages and adds language-specific adapters [19]. Whisper is an encoder-decoder Transformer trained on 680,000 hours of weakly supervised audio [20]. Wav2Vec2-BERT (w2v-BERT 2.0) combines contrastive learning with masked prediction in a Conformer encoder [6, 24].
Several studies adapt these models to Nepali. Rijal et al. fine-tuned Whisper on a curated dataset. On FLEURS, WER fell from 69.5% to 36.2% (small) and from 54.4% to 23.8% (medium) [23]. Pantha et al. applied LoRA and DoRA to a 1.55B-parameter Whisper model already fine-tuned on OSLR54, achieving relative CER improvements of 49.50% and 50.03%, respectively, while updating only 0.26% of the total parameters [14]. Ghimire et al. selected training samples for MMS adaptation with active learning and reached 6.77% CER [9]. Paudel and Sayami fine-tuned six pretrained models on OSLR54 and tested them on OpenSLR, FLEURS, and Common Voice. Their best result is 14.27% WER and 3.15% CER [17]. These studies use different test sets, so their numbers are not directly comparable.

2.4 Research Gaps

Three gaps remain. First, the reviewed Nepali ASR studies evaluate on read-speech test sets. Second, prior work reports mainly CER and WER. These metrics do not show where or why models fail. Third, to the authors' knowledge, no prior study has fine-tuned Wav2Vec2-BERT for Nepali or compared it with Whisper and MMS. This paper addresses all three gaps.

Table 1. : Selected public speech corpora containing Nepali, compared with *ShrutiParichhya* .

Corpus	Speech style
OSLR43 [25]	Read
OSLR54 [13]	Read
Common Voice [1]	Read
FLEURS [8]	Read
ShrutiParichhya (this work)	Natural

3. DATASET

This study uses two datasets. OSLR54 is used to fine-tune the models. *ShrutiParichhya* is used only to evaluate them. The two datasets differ in speech style, recording conditions, and transcription convention. Table 1 places both among public Nepali speech corpora.

3.1 Fine-Tuning Dataset: OSLR54

OSLR54 is a public Nepali ASR corpus distributed through OpenSLR [13]. It contains about 157,000 crowd-sourced utterances from 527 speakers. The total duration is about 165 hours. All recordings are mono, 16-bit linear PCM at 16 kHz, stored in FLAC format. Leading and trailing silence was trimmed from every utterance. The corpus was then shuffled with a fixed random seed of 42. Of the utterances, 90% were used for training and 10% (about 15,700 utterances) for validation. Speakers were not separated between splits: 526 of the 527 speakers appear in both the training and validation sets. The OSLR54 scores are therefore optimistic and not directly comparable with speaker-independent results in prior work. OSLR54 transcripts write numerals as digits (e.g., (ENG: 2046)) and not as words. All three models learned this convention during fine-tuning.

3.2 Evaluation Dataset: *ShrutiParichhya*

ShrutiParichhya is a new Nepali speech dataset for ASR evaluation. It was built to test models on speech that differs from OSLR54 in four ways: (1) the speech has a natural speaking style, (2) the speakers and recording conditions vary, (3) every utterance has an SNR value, and (4) the transcripts follow one consistent convention, including for numerals.

3.2.1 Collection. The audio was extracted from Nepali YouTube videos released under the Creative Commons Attribution (CC BY) licence. Videos were selected to cover both male and female speakers and different recording conditions. The videos come from many channels. Most cover politics, stories, and social information, and only a few are vlog-style. Most of the speech is spontaneously uttered with a natural speaking style. Recordings of female speakers were harder to find than recordings of male speakers. Some speakers were therefore also recorded by the authors, with their consent. The dataset still has fewer female speakers. The audio segments were first created automatically with voice activity detection and then inspected manually to ensure that no segment cuts through speech.

3.2.2 Transcription. Annotators transcribed every utterance manually in Devanagari script. All annotators are native Nepali speakers with formal education in Nepali

language. Annotation had two steps. First, each annotator drafted the transcript and checked spelling and grammar. Second, the annotators cross-verified each other's transcripts.
No expert linguist reviewed the transcripts. This is a limitation. The transcripts follow one convention for numerals. Numbers are written as words (for example, digit: into word: (ENG: 2046 → Two thousand forty six)). Punctuation was not transcribed. English words used in speech were transliterated into Devanagari.

3.2.3 Statistics. Table 2 gives the speaker and text statistics. The dataset has 4,428 utterances and about 5 hours of audio from 36 unique speakers (25 male and 11 female). The mean utterance duration is 4.06 s. The transcripts contain 37,809 words in total, of which 9,452 are unique.

Table 2. : Metadata of *ShrutiParichhya*

Description	Details
Total utterances	4,428
Total duration	∼5 hours
Utterance duration range	0.72–8.40 s
Unique speakers	36
Male Utterances	66%
Female Utterances	34%
Total words	37,809
Mean words per utterance	8.54
Vocabulary size	9,452
Mean utterance duration	4.06 s

3.2.4 SNR Annotation. The SNR of every utterance was estimated. SNR was computed in decibels as the logarithmic ratio between signal and noise power extracted via Librosa. The 90th and 10th percentiles of short-time RMS energy served as the signal and noise energy estimates, respectively. The utterances were grouped into four tiers: Clean (>30 dB), Good (20–30 dB), Moderate (10–20 dB), and Noisy (≤10 dB). The thresholds were chosen empirically rather than derived from acoustic theory, but they provide a practical scale for measuring model robustness across recording conditions. Table 3 gives the SNR distribution. SNR values range from 5.32 to 63.95 dB. The mean is 24.76 dB and the median is 23.34 dB. Most utterances fall in the Good tier (47.7%). The Noisy tier has only 35 utterances (0.8%).

Table 3. : SNR tier distribution of the evaluation dataset.

SNR Tier	Range	Count	Percentage
Clean	>30 dB	948	21.4%
Good	20–30 dB	2,114	47.7%
Moderate	10–20 dB	1,331	30.1%
Noisy	≤10 dB	35	0.8%
Total	5.32–63.95 dB	4,428	100%

3.2.5 Ethics and Availability. Most audio comes from public YouTube videos released under the CC BY license. The rest was recorded by the authors with the speakers' consent. The dataset are released [1] along with this paper, with attribution to the original creators. Speaker labels are left anonymous. Creators may request removal by contacting the authors.

[1] *ShrutiParichhya* : `https://github.com/rupakraj/ShrutiParichhya`

4. FINE-TUNING MULTILINGUAL MODELS FOR NEPALI ASR

Three pretrained multilingual speech models were fine-tuned on the OSLR54 corpus. Training was performed on a single NVIDIA GeForce RTX 4060 Ti GPU with 16 GB VRAM. Whisper-Small and Wav2Vec2-BERT were fully fine-tuned. MMS-1b-all was adapted using its standard adapter-based strategy, following the model's designed fine-tuning approach.
Table 4 reports WER and CER on the OSLR54 validation split. Wav2Vec2-BERT [6, 24] performs best (7.91% WER, 1.48% CER). Whisper-Small [20] is second (11.39% WER, 3.05% CER). MMS-1b-all [19] performs worst (45.49% WER, 11.57% CER). Wav2Vec2-BERT reaches its best WER at epoch 40. Its validation loss starts to diverge from the training loss at about epoch 12, which signals the onset of overfitting. The validation WER still improves until the end of training.
Whisper-Small also reaches its best WER at epoch 40. Its validation loss shows overfitting from epoch 5. The validation WER decreases more slowly than that of Wav2Vec2-BERT.

Table 4. : Fine-tuning results of the three pretrained models on the OSLR54 validation split.

Pre-trained Model	Hugging
Whisper-Small	ope
Wav2Vec2-BERT	face
MMS-1b-all	fa

Maximum 40 epochs with early stopping (MMS-1b-all stopped at epoch 14);

MMS-1b-all has the largest parameter count and the broadest multilingual pretraining. Yet it performs worst. Its validation loss and WER both plateau from about epoch 5, and early stopping ends training at epoch 14. Unlike the other two models, MMS-1b-all shows no divergence between training and validation loss. Its backbone was frozen, and only the language adapters and the CTC output head [11] were trained. The plateau is therefore consistent with limited trainable capacity rather than overfitting. This explanation was not tested with full fine-tuning.
These results measure in-distribution performance on OSLR54 read speech. Section 5 evaluates the same models on out-of-distribution speech.

5. EVALUATION ON THE *ShrutiParichhya*

5.1 Evaluation Protocol

The Three fine-tuned models were evaluated on *ShrutiParichhya* . Tables use WS, W2V, and MMS for Whisper-Small, Wav2Vec2-BERT, and MMS-1b-all respectively. No model was trained or tuned on *ShrutiParichhya* , and no external language model was used. Greedy decoding was used for all three models.
WER and CER are reported, computed as $(S + D + I)/N$. Here S, D, and I are the numbers of substituted, deleted, and inserted units, and N is the number of reference words (WER) or characters (CER). The median per-utterance

CER and the number of utterances with CER = 0 are also reported. The error counts are computed with jiwer[2].

5.2 Overall Results

Table 5 presents corpus-level WER and CER for all three models on *ShrutiParichhya* . Wav2Vec2-BERT achieves the best CER at 20.66%, closely followed by MMS-1b-all. The WER rankings reveal a different ordering: Wav2Vec2-BERT remains the best at 52.69%, followed by Whisper-Small, while MMS-1b-all performs significantly worse at 67.23% despite its competitive CER. Whisper-Small produces the most perfect transcriptions (CER = 0), with 97 utterances. All three models show substantially higher error rates than on the OSLR54 validation split, which reflects the greater difficulty of the out-of-distribution evaluation set. The difference in numeral convention also contributes to these error rates. Section 6.3 quantifies this effect.

Table 5. : Overall evaluation results on *ShrutiParichhya* - a new ASR Evaluation Dataset.

Model	WER(%)	CER(%)	Median CER(%)	CER=0
W2V	**52.69**	**20.66**	**18.42**	83
WS	57.41	30.56	26.87	97
MMS	67.23	20.92	18.46	40

5.3 Robustness Analysis

5.3.1 Effect of Acoustic Conditions (SNR). To measure the robustness of fine-tuned Nepali ASR models across a range of recording conditions, the models were evaluated with respect to the per-utterance SNR annotations. Table 6 reports CER and WER stratified by SNR tier. Several important findings emerge from this analysis.

Table 6. : Corpus-level CER(%) and WER(%) by SNR category.

	WS		W2V		MMS	
SNR	CER	WER	CER	WER	CER	WER
Clean (>30 dB)	31.7	58.2	22.3	**54.3**	**21.8**	66.2
Good (20–30 dB)	29.8	57.0	**19.8**	**52.1**	19.9	67.5
Moderate (10–20 dB)	30.5	57.0	**20.8**	**52.4**	21.5	67.5
Noisy (≤10 dB)	56.3	81.4	**27.2**	**60.1**	33.0	84.1

Clean audio is not the easiest tier. For all three models, the Clean tier has a higher CER than the Good and Moderate tiers. The Noisy tier holds only 35 utterances, so its values are unstable.

Whisper-Small degrades sharply under noise. Moving from Clean to Noisy conditions, Whisper-Small's CER rises from 31.7% to 56.3%, and its WER rises from 58.2% to 81.4%. This is the largest degradation observed across all models and SNR tiers. This tier holds only 35 utterances, so the result is tentative.

Wav2Vec2-BERT and MMS-1b-all degrade less under noise. Wav2Vec2-BERT shows a 4.9 point CER increase from Clean to Noisy conditions (22.3% → 27.2%), while MMS-1b-all shows an 11.2 point increase (21.8% → 33.0%).

Within the Clean-to-Moderate range, error rates change little. Wav2Vec2-BERT's CER fluctuates by at most 2.5 points between Clean and Moderate conditions (19.8%–22.3%), and MMS-1b-all's by at most 1.9 points (19.9%–21.8%).

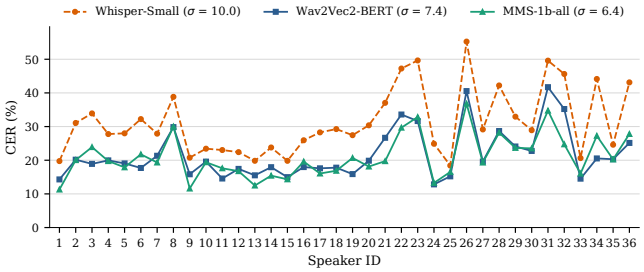

Fig. 1: Per-speaker CER comparison across all the evaluation speakers

5.3.2 Speaker-Level Variation. Figure 1 shows large differences between speakers. For every model, the CER of the hardest speaker is about three times that of the easiest speaker. The same speakers are hard for all three models: Speaker_26 and Speaker_31 are among the three hardest speakers for every model, with Speaker_26 the hardest for Whisper-Small and MMS-1b-all and Speaker_31 the hardest for Wav2Vec2-BERT. The difficulty therefore comes from the speaker or the recording, not from one model. Both speakers have a regional accent and a conversational style. The conversational style is consistent with the register mismatch discussed in Section 7.

5.4 Error Profile

5.4.1 Word-Level Error Types (S/D/I). Figure 2 shows the share of substitutions, deletions, and insertions in each model's word errors. Substitutions are the largest category for all three models (53.7–59.3%), followed by deletions (38.9–40.6%). Insertions differ most between models: 7.4% of Whisper-Small's errors, against 1.7% for Wav2Vec2-BERT and 0.7% for MMS-1b-all. This is consistent with Whisper's decoder entering repetitive loops under acoustic ambiguity. The CTC models produce far fewer insertions.

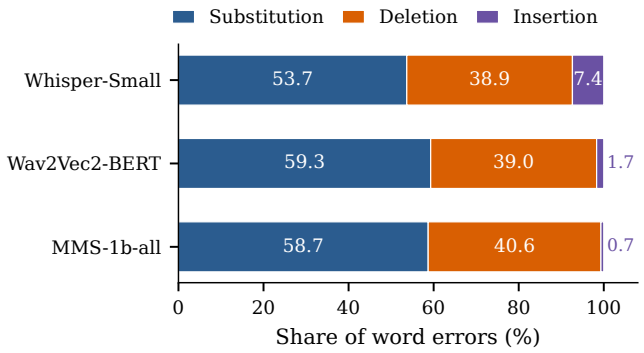

Fig. 2: Decomposition of word-level error metrics into substitutions, deletions, and insertions across the three fine-tuned models.

[2] https://github.com/jitsi/jiwer

5.4.2 Utterance-Level Distribution and Catastrophic Failures. Figure 3 shows that Whisper-Small is the least predictable model. Half of its utterances have a CER below 26.5%, but the distribution has a long right tail. The maximum CER is 1,112.5%, and 45 utterances (1.0%) exceed 100%. The standard deviation is 32.81 points. Wav2Vec2-BERT and MMS-1b-all are more stable (standard deviations 14.75 and 14.70 points). Only 6 and 4 of their utterances exceed 100% CER, and their maximum CERs are 183.3% and 142.9%.

Most of Whisper-Small's worst errors are hallucination loops, where the decoder repeats a syllable (Table 7). These loops are not limited to noisy audio. The utterance with the highest CER has a good SNR of 27.87 dB. The worst errors of the two CTC models occur on very short utterances, which may give too little context for alignment. One MMS-1b-all output was `NaN`, and its cause could not be traced.

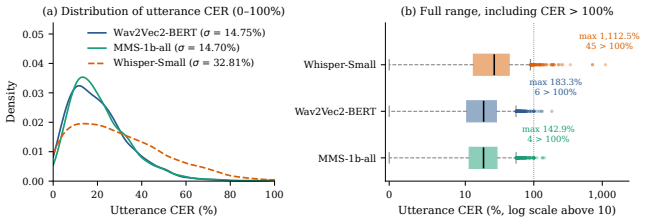

Fig. 3: Utterance-level CER distribution density profiles. The main panel shows the spread of typical performance; the inset log-scale view shows Whisper-Small's heavy tail extending past 100% CER. (WS = Whisper-Small, W2V = Wav2Vec2-BERT, MMS = MMS-1b-all)

Table 7. : Model outputs on the utterance producing Whisper-Small's highest recorded CER due to hallucination. Reference: (SNR 27.87 dB).

Model	Hypothesis	WER(%)	CER(%)
W2V		83.3	22.5
WS		1,733.3	1,112.5
MMS		100.0	20.0

6. LINGUISTIC ERROR ANALYSIS

WER and CER show how often a model fails, but not why. This section therefore analyses the errors by linguistic category. The ten categories were chosen from the most frequent error patterns in the alignments. Most of these patterns match known differences between Nepali spelling and pronunciation. Nepali spelling keeps Sanskrit-derived distinctions that speakers often do not pronounce [3]. Casual speech neutralises vowel length and consonant contrasts, and it simplifies conjunct consonants.

Character-level errors were extracted by aligning each hypothesis with its reference using Levenshtein alignment. The tables report counts over all 4,428 utterances of *ShrutiParichhya* . In the tables, WS, W2V, and MMS stand for Whisper-Small, Wav2Vec2-BERT, and MMS-1b-all. Section 6.1 covers errors inside words. Section 6.2 covers word boundaries and short function words. Section 6.3 covers the numeral convention mismatch, which is a transcription difference and not a model error.

Table 8. : Character-level confusion counts on *ShrutiParichhya* . An arrow (→) gives one direction (reference → hypothesis). A double arrow (↔) gives the sum of both directions. Bold marks the highest count in a row. WS = Whisper-Small, W2V = Wav2Vec2-BERT, MMS = MMS-1b-all.

Substitution	WS	W2V	MMS
Vowel length: Confusion on vowel markers and vowel (/)			
→	878	1,000	**1,012**
→	278	265	**353**
→	208	182	**392**
→	131	62	**159**
→	31	37	**61**
→	48	**57**	35
→	**15**	6	8
Nasalization			
→	10	16	**64**
→	0	0	**2**
→	3	1	**14**
→	23	**31**	23
∅ → (inserted)	372	**398**	254
Ba/va			
→	172	157	**291**
→	65	47	**175**
Sibilants			
→	112	112	**164**
→	78	51	**161**
→	41	38	**90**
→	23	6	**25**
→	11	14	**24**
→	8	5	**17**
Aspiration			
↔	96	57	**156**
↔	70	49	**167**
↔	72	40	**95**
↔	30	28	**63**
↔	58	38	**74**
Total	326	212	**555**
Retroflex/dental			
↔	83	75	**140**
↔	**56**	35	31
↔	**35**	6	11

6.1 Character-Level Errors

6.1.1 Vowel Length (Matra). One of the most frequent error types across all three models is confusion between short and long vowel matras—the diacritical vowel modifiers used in the Devanagari script. Table 8 reports the character substitution counts.

The short-u to long-u substitution (→) is the most frequent single substitution across all three systems (878, 1,000, and 1,012 instances). MMS-1b-all shows the highest counts in both directions of the / pair (392 and 353, compared to at most 278 for the other two models), suggesting a specific difficulty disambiguating this contrast rather than a one-directional bias.

6.1.2 Nasalization (Anusvara and Chandrabindu). The anusvara () and chandrabindu ()—the primary nasalization markers in Devanagari—exhibit distinct error profiles across the evaluated models. Table 8 reports the substitution counts.

Whisper-Small and Wav2Vec2-BERT's dominant substitution is anusvara mutating directly into chandrabindu (23 and 31 instances). MMS-1b-all instead favors materializing the anusvara as an explicit velar nasal (64 instances) or labial nasal (14 instances); this split does not track cleanly with model architecture, since Wav2Vec2-BERT and MMS-1b-all share the same non-autoregressive decoding style yet diverge here.

The most widespread error across all three models, however, is over-insertion of the chandrabindu () where no nasalization marker exists in the reference: 372 instances for Whisper-Small, 398 for Wav2Vec2-BERT, and 254 for MMS-1b-all—far exceeding any single substitution pattern above. This is consistent with the models conflating general vocalic resonance with active nasalization, though the acoustic cause was not directly tested here.

6.1.3 Consonant Contrasts. This subsection covers four consonant contrasts: ba/va, the sibilants, aspirated and unaspirated stops, and retroflex and dental consonants.

6.1.3.1 (ba/va) confusion.. Another pervasive consonant-level error is bidirectional confusion between [ba] and [va/wa]. Table 8 reports the substitution counts.
This confusion is phonetically motivated: in modern spoken Nepali, /b/ and /v/ are frequently realized as the same bilabial fricative [β] in word-medial or word-final position, though standard orthography preserves the distinction. MMS-1b-all shows the highest total confusion (466 instances), driven primarily by substitutions (291), roughly 1.7 times as frequent as the reverse (175).

6.1.3.2 Sibilant confusion (/ /).. All three models show systematic difficulty distinguishing the three Devanagari sibilants: (palatal, "sha"), (retroflex, also commonly pronounced "sha" in casual speech), and (dental, "sa"). Table 8 reports the substitution counts.
In colloquial spoken Nepali, these orthographic distinctions are frequently not realized acoustically, making them primarily a lexical rather than acoustic matter. The data shows this most strongly for the / pair, which accounts for the majority of sibilant confusions across all three models, while confusions involving are comparatively less frequent. MMS-1b-all shows the highest total confusion across all six substitution pairs, with a near-symmetric / split (164 vs. 161 instances).

6.1.3.3 Aspirated and unaspirated consonant confusion.. Nepali makes a systematic phonemic distinction between aspirated and unaspirated stops across all places of articulation. Table 8 reports the bidirectional confusion counts.
MMS-1b-all shows the highest total aspiration confusion (555 instances), roughly 2.6 times Wav2Vec2-BERT's total (212). The / pair is close to symmetric for MMS-1b-all (: 86; : 81), suggesting the model has little directional bias for this specific pair, though the source of the confusion itself was not directly tested. Aspiration is a strictly contrastive phonemic feature in Nepali—distinguishing lexemes such as (donation) from (paddy)—so these errors produce invalid lexical items even when surrounding characters are decoded correctly. Wav2Vec2-BERT shows the lowest total confusion (212 instances) among the three models.

6.1.3.4 Retroflex versus dental consonant confusion.. Nepali maintains a systematic phonemic contrast between retroflex consonants (, , , ,) and their dental counterparts (, , , ,). Table 8 reports the bidirectional confusion counts.
The / nasal pair accounts for the most confusions of the three, with MMS-1b showing nearly double the count of the other two models (140 vs. 83 and 75); the retroflex nasal is comparatively rare in colloquial Nepali and frequently merges with in casual speech. The direction of this confusion differs by model: MMS-1b-all substitutes more often than the reverse (80 vs. 60), while Whisper-Small and Wav2Vec2-BERT both show the opposite skew, overgenerating from inputs (49/34 and 41/34 respectively).
For the stop pairs, MMS-1b-all has the fewest / confusions (31) and Wav2Vec2-BERT the fewest / confusions (6), while Whisper-Small has the most for both pairs (56 and 35).

Table 9. : Halant () error counts.

Error Type	WS	W2V	MMS
Halant deleted	3,098	2,265	2,778
Halant inserted	978	463	386
Halant substituted	1,813	1,043	1,317
Delete-to-insert ratio	3.2:1	4.9:1	7.2:1

Table 10. : Word boundary error counts.

Error Type	WS	W2V	MMS
Spaces deleted (words merged)	8,783	8,141	10,656
Spaces inserted (words split)	2,153	665	509
Merge-to-split ratio	4.1:1	12.2:1	20.9:1

6.1.4 Conjunct Consonants (Halant). The halant () is the diacritical stroke that suppresses the inherent vowel of a consonant, joining separate characters into a unified conjunct cluster (e.g., + + =). It is one of the most structurally vital yet acoustically opaque components of Devanagari orthography. Table 9 summarizes the error profile across all three models.
All models delete far more halants than they insert, indicating a consistent bias toward breaking apart conjunct consonants rather than generating spurious ones. Dense multi-consonant clusters such as (competition), (to manage), and (self-reliant) are frequently simplified during inference, with the model dropping the virama and retaining only a subset of the surrounding consonants. Representative failure cases illustrate this pattern:

—REF:
HYP: (MMS-1b-all)
—REF:
HYP: (Wav2Vec2-BERT)
—REF:
HYP: (Whisper-Small)

In the final example, Whisper-Small collapses a five-word reference containing multiple conjunct clusters into a single short fragment, losing the majority of the underlying lexical content. The delete-to-insert ratio increases from Whisper-Small (3.2:1) to Wav2Vec2-BERT (4.9:1) to MMS-1b-all (7.2:1), consistent with—though not conclusively explained by—the shift from autoregressive to non-autoregressive decoding; other factors, such as MMS-1b-all's distinct tokenizer and vocabulary construction, may also contribute.

6.2 Word-Level Structural Errors

6.2.1 Word Boundary and Segmentation. Analysis of space character errors reveals a structural difference between the three models, as shown in Table 10.
All three models merge far more words than they split. MMS-1b-all is the most extreme, with a merge-to-split ratio of 20.9:1—10,656 missing spaces against only 509 inserted. This directly explains MMS-1b-all's anomalously high WER (67.23%) despite its competitive CER (20.92%): the model transcribes characters reasonably well, but systematically fails to place word boundaries, and each merged word pair

Table 11. : Mean CER (%) by reference word length (in characters).

Word Length (chars)	WS	W2V	MMS
1	113.1	130.9	240.6
2	84.7	99.7	142.1
3	57.9	60.9	98.3
4	47.9	48.0	72.9
5	44.6	43.4	63.5
6	38.2	36.1	54.2
7+	35.1	31.6	49.5

counts as two distinct word errors under standard WER scoring, severely inflating the metric.
Whisper-Small shows the most balanced ratio (4.1:1) of the three models, but also the highest number of false space insertions (2,153), consistent with its elevated insertion-error profile reported in Section *5.4.1*.
Table 6 shows that MMS-1b-all's WER stays high at every SNR level. Even under clean conditions, MMS-1b-all records a WER of 66.4%, compared to 54.3% for Wav2Vec2-BERT and 58.2% for Whisper-Small. This points to a structural word boundary issue rather than noise sensitivity.

6.2.2 Short Function Words. Character Error Rate stratified by reference word length reveals a consistent pattern across all three models: performance degrades sharply as word length decreases. Table 11 reports the stratified CER. Single-character words exceed 100% CER for all three models (peaking at 240.6% for MMS-1b-all), since even a small number of edits can exceed the length of a one-character reference. Error rates decrease monotonically with word length, reaching their lowest values for 7+ character words (31.6%–49.5% across models).
This pattern is linguistically expected: longer words offer a larger acoustic envelope and are more distinctly identifiable, while short Nepali words—postpositions (, , ,), determiners (,), and conjunctions (,)—are frequently reduced, rapidly articulated, and co-articulated with adjacent words in natural speech, making them harder to recover reliably.
MMS-1b-all's especially severe short-word CER (142.1%–240.6% for lengths 1–2) is consistent with the word-merging pattern in Section *6.2.1*: when spaces are dropped, short function words are absorbed into neighboring words and are not recovered as separate tokens.

6.3 Transcription Convention Mismatch (Numerals)

A systematic error observed across all three architectures is the transcription of verbalized number phrases as compact numeric digits instead of words. This pattern is prevalent across both the autoregressive Whisper-Small and the two CTC models (Table 12). This pattern stems from a convention mismatch rather than an acoustic or architectural failure: OSLR54's original transcripts represent numerals in digit form (e.g.,) rather than fully verbalized words, and all three models were fine-tuned on this convention. Since the reference transcripts of the evaluation dataset are fully verbalized, any digit-form output is penalized under standard CER/WER scoring even when the underlying number is transcribed correctly:

Table 12. : Cross-model numeral normalization frequencies.

Metric Type	WS	W2V	MMS
Sentences with numeral insertions	371	404	276
Total digit characters produced	1,165	1,067	663

—REF: HYP:
—REF: HYP:

Wav2Vec2-BERT produces this pattern in the most sentences (404), while Whisper-Small produces the most total digit characters overall (1,165). This inflates CER across all three models independent of genuine transcription quality, and highlights numeral verbalization as a training-corpus convention that future Nepali ASR work should standardize before fine-tuning.

7. DISCUSSION

Summary of Error Patterns: Table 13 lists the error categories explained in Section 6. It summarizes the type of error, the level of each model, and the likely cause. The levels (Lowest, Medium, High, Highest) are relative ordinal rankings assigned based on each model's error frequency and performance comparison within each specific category.
Speech Register Mismatch versus Model Capacity: The errors in this analysis are not random. They concentrate in specific places where natural Nepali speech differs from the OSLR54 training data. OSLR54 contains read speech with formal, standard spelling, whereas *ShrutiParichhya* contains spontaneous speech. The differences appear in four areas: colloquial phonemic mergers, simplified conjunct clusters, co-articulated function words, and numeral notation. These patterns suggest that the mismatch in speech register between training and test data is the main cause of the errors. They do not point to a limitation of any single model architecture.
Three observations support this reading.

(1) The gap between the two test sets is larger than the gap between the models. Whisper-Small and Wav2Vec2-BERT lose about 46 and 45 WER points from the OSLR54 validation split to *ShrutiParichhya* (Tables 4 and 5). MMS-1b-all loses about 22 points, because its OSLR54 WER is already high. The WER spread among the three models on *ShrutiParichhya* is 14.5 points.
(2) Recording noise does not explain the gap. Error rates change little across the Clean, Good, and Moderate tiers (Section *5.3.1*), and these tiers hold 99.2% of the utterances (Table 3).
(3) The errors concentrate in categories that arise from colloquial pronunciation and spelling conventions, and all three models show them (Table 13). The models differ mainly in how often they make these errors.

These observations are indirect. *ShrutiParichhya* differs from OSLR54 in speech style, speakers, recording conditions, and numeral convention at the same time. The gap therefore cannot be attributed to register alone. A direct test needs models fine-tuned on natural speech, which were not trained in this work. MMS-1b-all is a further case: its adapter-only fine-tuning limits its trainable capacity (Section 4), so its results do not separate capacity from register. The findings have three implications for Nepali ASR. First, read-speech test sets alone can overstate accuracy on natural speech, so evaluation should also include natural speech, as in *ShrutiParichhya* . Second, corpora should use one numeral

Table 13. : Linguistic error category summary.

Error Category	WS	W2V	MMS	
Short/long vowel confusion	High	High	Highest	
/ confusion	Medium	Medium	Highest	
Sibilant confusion (//)	Medium	Medium	Highest	
Anusvara → substitution	Medium	High	Highest	
Halant/conjunct deletion	Highest	Medium	High	
Aspiration confusion	Medium	Lowest	Highest	
Retroflex/dental confusion	High	Lowest	Highest	
Word merging	High	Medium	Highest	
Short word failurev	High	High	Highest	
Numeral normalization	High	Highest	Medium	

convention, and the convention should be checked before fine-tuning. Third, an external language model may reduce spelling errors that the acoustics cannot resolve, such as the sibilant and vowel-length confusions.

8. CONCLUSION

This paper fine-tuned three pretrained multilingual models, namely Wav2Vec2-BERT, MMS-1b-all, and Whisper-Small, on OSLR54 for Nepali ASR and evaluated them on *ShrutiParichhya* . Wav2Vec2-BERT performed best on both the OSLR54 validation split (7.91% WER, 1.48% CER) and *ShrutiParichhya* (52.69% WER, 20.66% CER). All three models degraded sharply on *ShrutiParichhya* . Error rates changed little across the Clean, Good, and Moderate SNR tiers, which hold 99.2% of the utterances, so recording noise does not explain the drop. The errors concentrated in colloquial phonemic mergers, deleted conjunct halants, merged word boundaries, and short function words. These patterns suggest that the mismatch in speech register between training and test data limits performance more than model capacity does. This evidence is indirect, because *ShrutiParichhya* differs from OSLR54 in several factors at once. The models also differed in failure mode. MMS-1b-all merged words most often (merge-to-split ratio 20.9:1), and Whisper-Small produced hallucination loops, with 45 utterances above 100% CER.

This work can be extended in several ways. Fine-tuning on natural Nepali speech would test the register explanation directly. Full fine-tuning of MMS-1b-all would separate trainable capacity from register. An external language model may reduce spelling errors that acoustics cannot resolve, such as sibilant and vowel-length confusions. Numeral normalisation before training would remove the digit-word mismatch. Finally, *ShrutiParichhya* can be extended with more variation of speakers, noisy recordings, and accent and code-mixing labels.

9. REFERENCES

[1] Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis Tyers, and Gregor Weber. Common voice: A massively-multilingual speech corpus. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4218–4222, Marseille, France, May 2020. European Language Resources Association.

[2] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Advances in Neural Information Processing Systems*, 2020.

[3] Bal Krishna Bal. *Structure of Nepali Grammar*. 2004.

[4] Elina Baral and Sagar Shrestha. Large vocabulary continuous speech recognition for nepali language. In *International Journal of Signal Processing Systems*, volume 8, pages 68–73, 2020.

[5] Bharat Bhatta, Basanta Joshi, and Ram Krishna Maharjhan. 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, Taipei, Taiwan, September 2020. The Association for Computational Linguistics and Chinese Language Processing (ACLCLP).

[6] Yu-An Chung, Yu Zhang, Wei Han, Chung-Cheng Chiu, James Qin, Ruoming Pang, and Yonghui Wu. w2v-BERT: Combining contrastive learning and masked language modeling for self-supervised speech pre-training. In *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 244–250. IEEE, 2021.

[7] Alexis Conneau, Alexei Baevski, Ronan Collobert, Abdelrahman Mohamed, and Michael Auli. Unsupervised cross-lingual representation learning for speech recognition, 2020.

[8] Alexis Conneau, Min Ma, Simran Khanuja, Yu Zhang, Vera Axelrod, Siddharth Dalmia, Jason Riesa, Clara Rivera, and Ankur Bapna. FLEURS: Few-shot learning evaluation of universal representations of speech. In *2022 IEEE Spoken Language Technology Workshop (SLT)*, pages 798–805. IEEE, 2023.

[9] Rupak Raj Ghimire, Bal Krishna Bal, and Prakash Poudyal. Active learning approach for fine-tuning pretrained 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, December 2023. NLP Association of India (NLPAI).

[10] Rupak Raj Ghimire, Bal Krishna Bal, Balaram Prasain, and Prakash Poudyal. 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, December 2023. NLP Association of India (NLPAI).

[11] Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd International Conference on Machine Learning (ICML)*, pages 369–376, 2006.

[12] Tahir Javed, Janki Nawale, Eldho George, and Others. IndicVoices: Towards building an inclusive multilingual speech dataset for Indian languages. In *Findings of the Association for Computational Linguistics: ACL 2024*, Bangkok, Thailand, 2024. Association for Computational Linguistics.

[13] Oddur Kjartansson, Supheakmungkol Sarin, Knot Pipatsrisawat, Martin Jansche, and Linne Ha. Crowd-sourced speech corpora for javanese, sundanese, sinhala, nepali, and bangladeshi bengali. In *Proc. The 6th Intl. Workshop on Spoken Language Technologies for Under-Resourced Languages*, pages 52–55, 2018.

[14] Kiran Pantha, Rupak Raj Ghimire, and Bal Krishna Bal. Speech personalization using parameter efficient fine-tuning for Nepali speakers. In *Proceedings of the Fifth Workshop on Language Technology for Equality, Diversity and Inclusion (LT-EDI 2025)*, pages 190–199. Association for Computational Linguistics, 2025.

[15] Basanta Joshi Paribesh Regmi, Arjun Dahal. Nepali speech recognition using rnn-ctc model. *International Journal of Computer Applications*, 178(31):1–6, Jul 2019.

[16] Shishir Paudel, Bal Krishna Bal, and Dhiraj Shrestha. 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, September 2023.

[17] Suman Paudel and Sarbin Sayami. Comparative analysis of multilingual pre-trained models for nepali automatic speech recognition. *arXiv preprint arXiv:2608.12327*, 2026.

[18] Jenny Poudel, Ankit Dahal, Rishikesh Kumar Sharma, Rupak Tiwari, Rupak Raj Ghimire, and Bal Krishna Bal. Nepconformer: A conformer-based nepali automatic speech recognition system. In *Computing and Machine Learning*, pages 167–178, Singapore, 2026. Springer Nature Singapore.

[19] Vineel Pratap, Andros Tjandra, Bowen Shi, Paden Tomasello, Arun Babu, Sayani Kundu, Ali Elkahky, Zhaoheng Ni, Apoorv Vyas, Maryam Fazel-Zarandi, Alexei Baevski, Yossi Adi, Xiaohui Zhang, Wei-Ning Hsu, Alexis Conneau, and Michael Auli. Scaling speech technology to 1,000+ languages. *J. Mach. Learn. Res.*, 25(1), January 2024.

[20] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org, 2023.

[21] Dan Raj Regmi. Languages of Nepal in the 2021 census: An overview. Central Department of Linguistics, Tribhuvan University. Available on Academia.edu, 2021. Accessed: 2026-09-19.

[22] Sunil Regmi and Bal Krishna Bal. 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, December 2021. NLP Association of India (NLPAI).

[23] Sanjay Rijal, Shital Adhikari, Manish Dahal, Manish Awale, and Vaghawan Ojha. Whisper finetuning on nepali language, 2024.

[24] Seamless Communication, Loïc Barrault, Yu-An Chung, Mariano Coria Meglioli, David Dale, Ning Dong, et al. Seamless: Multilingual expressive and streaming speech translation. *arXiv preprint arXiv:2312.05187*, 2023.

[25] Keshan Sodimana, Knot Pipatsrisawat, Linne Ha, Martin Jansche, Oddur Kjartansson, Pasindu De Silva, and Supheakmungkol Sarin. A Step-by-Step Process for Building TTS Voices Using Open Source Data and Framework for Bangla, Javanese, Khmer, Nepali, Sinhala, and Sundanese. In *Proc. The 6th Intl. Workshop on Spoken Language Technologies for Under-Resourced Languages (SLTU)*, pages 66–70, Gurugram, India, August 2018.

[26] Manish K. Ssarma, Avaas Gajurel, Anup Pokhrel, and Basanta Joshi. Hmm based isolated word nepali speech recognition. In *2017 International Conference on Machine Learning and Cybernetics (ICMLC)*, volume 1, pages 71–76, 2017.

[27] Yogendra P. Yadava, Andrew Hardie, Ram Raj Lohani, Bhim N. Regmi, Srishtee Gurung, Amar Gurung, Tony McEnery, Jens Allwood, and Pat Hall. Construction and annotation of a corpus of contemporary nepali. *Corpora*, 3(2):213–225, 2008.