Everest Advances in Science and Technology (EAST), Vol. 2, No. 1, 2026
ISSN: 3102-0410 (Print), 3102-0429 (Online)
DOI: 10.3126/east.v2i1.98619

A Nepali-Accented English Evaluation Dataset for Automatic Speech Recognition

SANTOSH DAHAL[1*], KIRAN CHANDRA DAHAL[2]

Abstract. Automatic speech recognition (ASR) systems perform strongly on native-English benchmarks, yet their accuracy degrades sharply when the input speech comes from under-represented non-native accents. Nepali-accented English is particularly under-served: existing resources either focus on native Nepali speech, cover broader multi-accent settings without dedicated Nepali evaluation, or provide only limited Nepali-accent coverage. This paper presents a Nepali-accented English evaluation dataset designed to support robust ASR benchmarking under accent mismatch. The corpus was collected through a web-based platform that did not collect directly identifying metadata and contains recordings from 57 speakers. Each session follows a fixed 22-prompt protocol consisting of 11 phonetic prompts, 10 domain prompts, and 1 spontaneous prompt, providing complementary coverage of pronunciation, topical vocabulary, and natural speaking style. In addition to transcribed speech, the dataset includes participant metadata for coarse exploratory subgroup analysis and speaker-level manual recording-quality labels. Manual quality assessment shows that 50.9% of sessions are clean and 42.1% contain only mild noise. As a descriptive reference, open-source ASR baselines are substantially worse on this corpus than the corresponding LibriSpeech test-clean values reported in official model cards, reaching 38.15–55.00% WER on the collected set versus reported 2–4% WER on LibriSpeech test-clean. These baseline results position the corpus as a practical held-out resource for evaluating accent robustness and out-of-distribution generalization on Nepali-accented English.

Keywords: automatic speech recognition, Nepali-accented English, accented speech, evaluation dataset, word error rate

[1,2] Department of Electronics and Computer Engineering, Institute of Engineering, Thapathali Campus, Tribhuvan University
E-mail: santosh.080msise16@tcioe.edu.np;
dahalkc@tcioe.edu.np
* Corresponding author

Manuscript received: 19 February 2026; Revised: 13 June 2026; Accepted: 15 July 2026.

1. Introduction

Recent ASR systems achieve very low error rates on standard native-English benchmarks, but these gains do not transfer reliably to accented speech. The gap is especially important for under-represented accents, where both training and evaluation resources remain scarce. Nepali-accented English belongs to this category: it appears only sparsely in existing corpora even though English is widely used in education, technology, public services, and professional communication in Nepal. The lack of a focused benchmark makes it difficult to answer a basic question: how well do current ASR models handle English spoken by Nepali L1 speakers? Multi-accent resources such as L2-ARCTIC provide valuable non-native English data, but they do not include Nepali speakers [1]. Other accented-English corpora cover a broader set of foreign accents yet still omit Nepali as a target evaluation accent [2]. Meanwhile, Svarah demonstrates the importance of evaluating ASR on Indian accents and includes a Nepali-accented English category, but the speakers in that subset are from India rather than Nepal. Its scope is also broader than a Nepali-specific benchmark and is not centered on controlled phonetic coverage for this accent [3]. Existing Nepali ASR work, in contrast, focuses on speech recognition for the Nepali language itself rather than English spoken with a Nepali accent [4, 5].

This paper presents a dataset created to fill that gap. The corpus is designed as an *evaluation-focused* benchmark rather than a large-scale pretraining resource. The collection protocol aims to keep per-speaker coverage comparable while still capturing three complementary dimensions of variation: phonetic coverage, domain vocabulary, and spontaneous speech. The resulting dataset is then used as a held-out benchmark for ASR evaluation.

The main contributions of this paper are as follows:

- A Nepali-accented English evaluation corpus with 57 speakers collected through a consistent prompt-based protocol.
- A prompt design that combines phonetic prompts, domain prompts, and spontaneous speech within every participant session.
- Metadata and speaker-level recording-quality annotations that support coarse exploratory subgroup analysis and data-quality filtering.
- Baseline evidence that modern open-source ASR systems degrade substantially on the corpus compared with their reported LibriSpeech performance.

2. Related Work

Several existing datasets motivate the need for a Nepali-specific accented-English benchmark. L2-ARCTIC is a widely used non-native English corpus containing six L1 accent groups and carefully read speech, making it valuable for accent adaptation research; however, it does not include Nepali-accented English [1]. The CSLU Foreign Accented English corpus broadens accent coverage, but it is likewise not targeted toward Nepali speakers [2]. Svarah extends the evaluation landscape by covering Indian-accented English and explicitly includes a Nepali-accented speech subset, showing that ASR systems experience pronounced degradation when moving away from native-English benchmarks. However, the speakers in that subset are from India rather than Nepal [3]. Still, Svarah is a multilingual-accent benchmark spanning many language backgrounds and speaking conditions; it does not offer a Nepali-only evaluation design with fixed per-speaker phonetic and topical coverage.

On the Nepali side, prior work has concentrated on ASR for native Nepali speech using convolutional, recurrent, and Conformer-style architectures [**?**, 4, 5]. These studies are important for low-resource speech technology in Nepal, but they address a different problem from recognizing English produced with Nepali phonetic transfer. The present corpus therefore complements both lines of work: it is not a replacement for broad multi-accent English datasets, and it is not a native Nepali ASR corpus; instead, it specifically targets evaluation of Nepali-accented English.

TABLE 1. Overview of the Nepali-accented English evaluation dataset.

Aspect	**Description**
Primary goal	Held-out evaluation benchmark for Nepali-accented English ASR
Participants	57 speakers
Collection mode	Remote recording through a web-based platform without directly identifying metadata
Speech styles	Controlled read speech and one spontaneous response per session
Session structure	22 prompts per participant
Prompt mix	11 phonetic prompts, 10 domain prompts, 1 spontaneous prompt
Metadata	Age range, gender, highest qualification, occupation, ethnicity, district of origin, and topic of interest
Quality annotation	Manual speaker-level recording-quality label on a five-point ordinal scale
Intended use	Accent-robust ASR evaluation, coarse subgroup analysis, and error analysis

TABLE 2. Prompt design used in each recording session.

Prompt type	**Count**	**Purpose**
Phonetic	11	Cover English phonemes and common phonotactic contexts using CMU Arctic-style prompts
Domain	10	Introduce topic-specific vocabulary from domains such as technology, science, history, environment, culture, health, education, geography, and society
Spontaneous	1	Elicit natural speech from a participant-selected interest topic such as sports, reading, music, or cooking

3. Dataset Design and Collection Protocol

The dataset was built around a simple design principle: each participant should contribute a comparable but diverse set of utterances so that the final corpus supports controlled evaluation rather than ad hoc sampling. Table 1 summarizes the overall design.

3.1. **Collection Platform and Workflow.** Data were collected through a dedicated single-page web application designed for guided remote participation. The participant workflow contains four stages: session initialization without directly identifying metadata, consent acknowledgement with metadata entry, prompt-wise recording with re-record or retry options, and final review with submission. This structure was chosen to keep the data collection experience lightweight while still improving consistency across speakers and sessions.
The platform records only analysis-oriented metadata needed for curation and descriptive analysis. Session information is stored together with timestamps and completion status, allowing the final evaluation set to be curated without collecting direct personal identifiers. Because speech recordings can still be identifying, any dataset distribution or reuse should remain within participant-consent boundaries and appropriate access control.

3.2. **Prompt Design.** Each participant session contains 22 prompts drawn from three prompt types, shown in Table 2. The phonetic prompts target pronunciation coverage, the domain prompts broaden lexical and topical variation, and the spontaneous prompt captures less constrained speech with more natural discourse structure.
The phonetic prompts were designed in the style of the CMU Arctic prompt inventory [6] and organized to keep phonetic coverage balanced across participants. Domain prompts were curated

Table 3. ASR back end used for spontaneous transcription.

Item	Value
Model	`gemini-3.1-pro-preview`
No-speech token	`[NO_SPEECH]` (exact)

from public text sources, then normalized for readability and collection consistency. The spontaneous prompt is tied to the participant's self-reported topic of interest so that the dataset includes at least one segment with less scripted speaking behavior per session.

3.3. **Transcription of Spontaneous Speech.** Spontaneous responses were derived from elicited recordings whose metadata designated the spontaneous condition, as distinct from read phonetic or domain-matched material. For each item, the corresponding waveform was assigned a short speaker identifier and a unique session–utterance ID, and stored under the `spontaneous` partition of the corpus.

Segmentation. Recordings longer than 20 s were split prior to automatic transcription. Silence boundaries were approximated with FFMPEG `silencedetect` using a noise threshold of −35 dB and a minimum silence duration of 1.0 s. Resulting speech intervals were further constrained so that no segment exceeded 20 s, with a minimum segment length of 0.25 s. Recordings of 20 s or less were left as a single segment.

Automatic speech recognition. Each segment was transcribed with the Google Generative Language API using the Python `google-genai` client. The generative model was `gemini-3.1-pro-preview`. For each call, the model received the audio in WAV form (`audio/wav`) together with a fixed text instruction. Whitespace in the model output was normalized; an empty hypothesis was treated as a no-speech case. Table 3 summarizes the ASR back end used for spontaneous transcription.

Instruction prompt (exact text). The following prompt was used for transcription:

```
Listen to this audio and transcribe the speech.
Rules:
- Output only the spoken words, in English (including any Nepali-accented
English as heard).
- No timestamps, no speaker labels, no ‘‘Transcript:’’ prefix, no
markdown.
- If there is no intelligible speech, output exactly: [NO_SPEECH]
```

Transcript assembly. For multi-segment recordings, per-segment transcriptions were concatenated in temporal order with inter-word spaces to form one line per session–utterance, stored in `trans.txt` in the form *utterance-id* followed by the full transcript, consistent with the read side of the corpus. Segments for which the model returned the token `[NO_SPEECH]` (exactly) were treated as non-contributing at the segment level when assembling the line.

3.4. **Metadata and Privacy.** The collection form supports age range (18–30, 30–45, 45–60, 60+), gender, highest qualification, occupation domain, ethnicity category, district of origin, and topic of interest. In the present paper, we report six summary attributes for analysis: gender, ethnicity, age group, highest qualification, occupation domain, and topic of interest. District of origin was collected during data acquisition but is not broken down in the aggregate tables below.

4. Corpus Statistics and Recording Quality

The final evaluation set contains 57 participants. Figure 1 shows the geographic distribution of participants by reported district of origin across Nepal, while Tables 4 and 5 summarize the collected metadata distributions.

The collected subset is not demographically balanced, but it does capture meaningful variation in age, education, ethnicity, occupation, and self-selected topics. The distribution is dominated by younger adults (78.9% in the 18–30 range), participants with bachelor's-level education (66.7%), and contributors from engineering or technology backgrounds (54.4%). Topic-of-interest responses

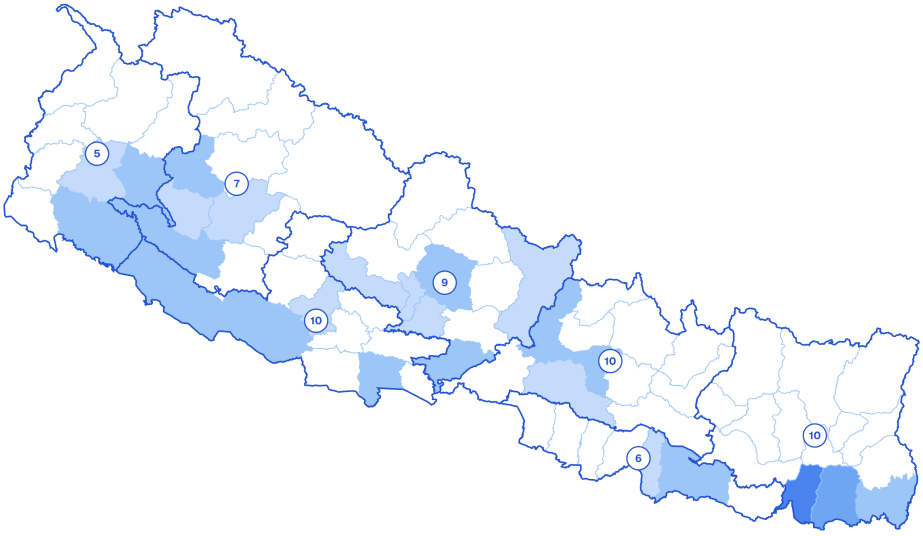

FIGURE 1. Geographic distribution of participants by reported district of origin across Nepal.

TABLE 4. Participant profile summary ($n = 57$).

Attribute	Category	Count	%
Gender	Male	37	64.9
Gender	Female	20	35.1
Age group	18–30	45	78.9
Age group	30–45	10	17.5
Age group	45–60	2	3.5
Qualification	Bachelors	38	66.7
Qualification	Masters	12	21.1
Qualification	+2 (higher secondary)	7	12.3

are led by travelling, sports, and reading, which helps diversify the spontaneous-speech portion of the corpus.

4.1. **Speaker-Level Recording Quality.** Because all utterances from a participant were recorded within a single session, ambient noise, microphone placement, and room acoustics are largely session-constant. For that reason, recording quality was summarized at the speaker level rather than the utterance level. The annotation procedure sampled three utterances uniformly at random for each speaker, played them under fixed listening conditions, and assigned one of five ordinal labels: clean, mild noise, noisy, very noisy, or bad recording.
Table 6 shows the resulting distribution. The large majority of sessions (93.0%) fall in the clean or mild-noise categories, suggesting that the dataset is generally suitable for ASR benchmarking while still preserving realistic variability from remote volunteer recordings. These labels should be interpreted as coarse speaker-level tags for exploratory filtering and error analysis rather than as a formally validated acoustic annotation layer.

5. Benchmark Evaluation with Open-Source ASR Models

To assess the difficulty of the corpus, the dataset was used as a held-out benchmark for several open-source ASR models. Table 7 contrasts the combined WER on the collected set with each model's reported LibriSpeech test-clean WER from the corresponding official model card [7, 8, 9, 10].

TABLE 5. Participant context attributes ($n = 57$).

Attribute	Category	Count	%
Ethnicity	Brahmin/Chhetri	43	75.4
Ethnicity	Madhesi	6	10.5
Ethnicity	Janajati	4	7.0
Ethnicity	Other	2	3.5
Ethnicity	Newar	1	1.8
Ethnicity	Dalit	1	1.8
Occupation	Engineering/Technology	31	54.4
Occupation	Student	13	22.8
Occupation	Health/Medicine	3	5.3
Occupation	Business/Management	3	5.3
Occupation	Education/Teaching	2	3.5
Occupation	Science	2	3.5
Occupation	Law/Government	2	3.5
Occupation	Other	1	1.8
Interest	Travelling	17	29.8
Interest	Sports	14	24.6
Interest	Reading	9	15.8
Interest	Movies/Entertainment	5	8.8
Interest	Painting/Drawing	4	7.0
Interest	Music	4	7.0
Interest	Cooking	3	5.3
Interest	Stitching/Sewing	1	1.8

TABLE 6. Manual speaker-level recording-quality labels.

Label	Count	%
Clean	29	50.9
Mild noise	24	42.1
Noisy	3	5.3
Very noisy	0	0.0
Bad recording	1	1.8

TABLE 7. Reported LibriSpeech test-clean WER from official model cards and observed WER on the collected Nepali-accented English set.

Model	LibriSpeech test-clean	Collected set
Conformer Small	3.70	46.00
Conformer Large	2.20	38.30
Whisper Small	3.43	55.00
Whisper Large	3.00	38.15

Because the LibriSpeech values are taken from official model cards rather than a single re-run pipeline, the comparison should be interpreted as a descriptive cross-dataset reference rather than a controlled estimate of accent-only degradation. For fully reproducible reuse, the benchmark release should package the corresponding reference transcripts, normalization rules, and scoring scripts alongside the audio and metadata.

All tested models degrade substantially on the collected benchmark. Even the strongest large-model baselines, Conformer Large and Whisper Large, rise from approximately 2–3% WER on LibriSpeech test-clean to about 38% WER on the Nepali-accented set. The smaller models degrade

even more sharply, with WER reaching 46.00% for Conformer Small and 55.00% for Whisper Small. These descriptive results reinforce the dataset's usefulness as a stress test for accent robustness and show that strong native-English benchmark performance does not necessarily imply reliable performance on Nepali-accented English under the present evaluation setup.

6. Discussion, Intended Use, and Limitations

The dataset has three main strengths. First, it combines controlled phonetic coverage with broader topical prompts and spontaneous speech, making it more informative than a read-only benchmark. Second, the metadata support coarse exploratory subgroup and robustness analysis beyond a single aggregate WER number, although several categories remain too small for strong comparative claims. Third, the recording-quality labels make it possible to analyze or filter the effect of session noise on ASR performance.

At the same time, the current corpus has clear limitations. The participant pool is skewed toward younger, highly educated speakers and is dominated by engineering or technology backgrounds. The recordings are also collected remotely rather than under studio conditions, which introduces device and environment variability even when most sessions remain usable. Finally, the dataset is best suited to held-out evaluation and error analysis; its size and collection design are not intended to replace broader multi-accent training corpora.

We therefore recommend the corpus for three use cases in particular. The first is benchmarking accent-robust ASR. The second is analyzing errors across prompt types or coarse participant groupings. The third is measuring out-of-distribution generalization when Nepali-accented English is absent from training data. We do not recommend using the metadata for speaker profiling or any ranking of participants; the metadata are included to support coarse exploratory subgroup and robustness analysis, not identity inference.

7. Conclusion

This paper presented a Nepali-accented English evaluation dataset for ASR containing recordings from 57 speakers collected through a structured web-based protocol. The dataset pairs fixed per-session prompt coverage with demographic and quality annotations, creating a practical benchmark for studying accent robustness. Baseline experiments with open-source Conformer and Whisper models show substantially higher WER on the collected set than the corresponding published LibriSpeech reference values, reinforcing the need for targeted evaluation on underrepresented accents. The corpus provides a useful foundation for future work on Nepali-accented English recognition, accent-aware adaptation, and more inclusive ASR evaluation.

The dataset is publicly available on GitHub at the project repository.

References

[1] G. Zhao, S. Sonsaat, A. Silpachai, I. Lucic, E. Chukharev-Hudilainen, J. Levis, and R. Gutierrez-Osuna, "L2-ARCTIC: A non-native English speech corpus," in *Proc. Interspeech 2018*, 2018, pp. 2783–2787.

[2] T. Lander. (2022) CSLU: Foreign accented English. Dataset, Release 1.2, doi: 10.5683/SP2/K7EQTE.

[3] T. Javed, K. S. Bhogale, A. Raman, P. Mukherjee, S. Cauligi, K. Thopalli, S. Ghosh, V. Seshadri, and M. M. Khapra, "Svarah: Evaluating English ASR systems on Indian accents," in *Proc. Interspeech 2023*, 2023, pp. 4064–4068.

[4] J. Banjara, K. R. Mishra, J. Rathi, K. Karki, and S. Shakya, "Nepali speech recognition using CNN and sequence models," in *Proc. IEEE Int. Conf. Mach. Learn. Appl. Netw. Technol. (ICMLANT)*, 2020, pp. 1–5.

[5] M. Dhakal *et al.*, "Automatic speech recognition for the Nepali language using CNN, bidirectional LSTM and ResNet," in *Proc. IEEE Int. Conf. Inventive Comput. Technol. (ICICT)*, 2022, pp. 515–521.

[6] J. Kominek and A. W. Black, "The CMU ARCTIC speech databases," Lang. Technol. Inst., Carnegie Mellon Univ., Pittsburgh, PA, USA, Tech. Rep. Tech. Rep. CMU-LTI-03-177, 2003.

[7] NVIDIA. (2024) NVIDIA Conformer-CTC Small (en-US). Hugging Face model card. [Online]. Available: https://huggingface.co/nvidia/stt_en_conformer_ctc_small

[8] ——. (2024) NVIDIA Conformer-CTC Large (en-US). Hugging Face model card. [Online]. Available: https://huggingface.co/nvidia/stt_en_conformer_ctc_large

[9] OpenAI. (2022) Whisper Small. Hugging Face model card. [Online]. Available: https://huggingface.co/openai/whisper-small

[10] ——. (2022) Whisper Large-v2. Hugging Face model card. [Online]. Available: https://huggingface.co/openai/whisper-large-v2