

AdiDhwani: Preservation-Oriented ASR for Endangered Bhili and Mundari Languages

Pooja Singh^{1*} M Kaab Bin Shahid^{2*} Sandeep Kumar^{1,3}

¹Indian Institute of Technology Delhi, New Delhi, India

²University of Stuttgart, Stuttgart, Germany

³PiByte AI, New Delhi, India

eez228470@ee.iitd.ac.in st200141@stud.uni-stuttgart.de

Abstract

India’s tribal (*Janjatiya/Adivasi*) languages remain absent from public ASR evaluation, despite their importance for oral history, ecological knowledge, and cultural continuity. We present AdiDhwani, a preservation-oriented ASR benchmark for Bhili (ISO 639-3: bhh), India’s largest non-scheduled language, and Mundari (ISO 639-3: unr), a Vulnerable North Munda language. The benchmark comprises 56.5 hours of community-validated Bhili speech and 11.7 hours of Mundari speech with fixed evaluation splits and reproducible baselines. We evaluate multilingual ASR under zero-shot transfer, target-language fine-tuning, donor-language initialization, synthetic supervision, and KenLM-assisted decoding. Target-language supervision substantially improves WER, yet lower edit distance does not always reflect better documentation fidelity: models replace attested Bhili forms with Hindi cognates or delete grammatically meaningful Mundari suffixes while still achieving competitive scores. We introduce Hindi-Normalization Bias (HNB) and its alignment-constrained variant (HNBA), Morphological Fragmentation Rate (MFR), and Suffix Deletion Rate (SDR) as preservation-oriented diagnostics beyond WER. Human transcriber baselines show that these gaps are not fixed limits. AdiDhwani provides the first controlled ASR evaluation framework for these two under-documented Indian languages and supports language-preserving speech recognition research in endangered-language settings.

1 Introduction

Automatic speech recognition (ASR) has advanced rapidly through multilingual pretraining, large-scale weak supervision, and shared evaluation benchmarks (Besacier et al., 2014; Pratap et al., 2024; Radford et al., 2023). Yet these

gains remain unevenly distributed. Languages with substantial digital resources continue to benefit most, while many tribal, Indigenous, and under-documented languages remain missing from public ASR evaluation (Panayotov et al., 2015; Wang et al., 2021; Joshi et al., 2020; Conneau et al., 2023). This absence is not only a matter of benchmark coverage. For many communities, speech recordings are primary records of oral history, ecological knowledge, ritual practice, and collective memory (Etman and Beex, 2015). Without reliable transcription, such recordings remain difficult to search, annotate, archive, or reuse for language documentation and revitalization (Cox, 2011; Bird, 2020; Nekoto et al., 2020).

Tribal Languages. We focus on two under-documented Indian tribal languages: Bhili and Mundari.¹ Bhili (ISO 639-3: bhh) is a Western Indo-Aryan language spoken across Gujarat, Rajasthan, Madhya Pradesh, and Maharashtra. With approximately 10.4 million speakers, it is India’s largest non-scheduled language.² Its sustained contact with Hindi, Gujarati, and Marathi makes cross-lingual transfer plausible, but also creates the risk that ASR systems normalize community-attested Bhili forms toward dominant Hindi variants. Mundari (ISO 639-3: unr) is a North Munda language of the Austroasiatic family and is classified as Vulnerable by UNESCO.³ Its agglutinative morphology, checked consonants, and multi-script usage differ sharply from the Indo-Aryan

¹Speaker counts are approximate 2011 Census lower-bound estimates; ISO 639-3 codes and language-family classifications follow Glottolog. Bhili Glottolog entry: <https://glottolog.org/resource/languoid/id/bhil1251>. Mundari Glottolog entry: <https://glottolog.org/resource/languoid/id/mund1320>.

²Bhili speaker-count reference: <https://en.wikipedia.org/wiki/Bhil>.

³Mundari UNESCO vulnerability source: <https://unesdoc.unesco.org/ark:/48223/pf0000397158.locale=en>.

*Equal contribution.

languages that dominate existing Indic ASR resources (Anderson, 2015; Osborne et al., 2012).

Evaluation Gap. WER and CER are useful for comparing recognition accuracy, but they do not distinguish ordinary transcription errors from errors that reduce documentary fidelity (Kim et al., 2021; Lignos and Marcus, 2013). A system may lower WER by replacing a Bhili form with a Hindi cognate, or by recognizing a Mundari stem while deleting suffixes that encode grammatical meaning. In both cases, transcript becomes less useful as a linguistic record even when edit distance improves.

Overview of AdiDhwani. To address these gaps, we introduce AdiDhwani, a paired ASR benchmark for Bhili and Mundari with 56.5 hours of quality-controlled Bhili speech and 11.7 hours of quality-controlled Mundari speech, collected through community-centered field workflows. To our knowledge, no existing resource provides evaluation-ready ASR splits, reproducible baselines, and preservation-oriented diagnostics for both languages (Table 1).

The benchmark provides fixed speaker-disjoint splits, a deterministic transcript-preprocessing pipeline applied identically to references and hypotheses, and reproducible baselines. The pipeline is language-preserving: it never rewrites community-attested Bhili or Mundari forms toward Hindi. We evaluate ASR models under five regimes: transfer, target-language fine-tuning, donor-language initialization, synthetic supervision, and KenLM-assisted decoding. Beyond WER and CER, we report human transcriber baselines, the CER/WER ratio, and four preservation-oriented diagnostics: Hindi-Normalization Bias (HNB), its alignment-constrained variant (HNBA), Morphological Fragmentation Rate (MFR), and Suffix Deletion Rate (SDR). These metrics measure whether ASR outputs preserve community-attested lexical and morphological forms rather than only reducing edit distance.

Contributions.

1. We introduce AdiDhwani, the first controlled ASR benchmark for Bhili and Mundari, with community-validated transcripts, fixed speaker-disjoint splits, and reproducible baselines.
2. We describe a documentation-oriented data workflow combining bilingual elicitation, au-

tomatic target–translation segmentation, and maker–checker–superchecker adjudication with community and expert review.

3. We provide a controlled evaluation across transfer, target-language fine-tuning, donor-language initialization, synthetic supervision, and KenLM-assisted decoding.
4. We introduce HNB, HNBA, MFR, and SDR as preservation-oriented diagnostics for documentation-relevant ASR failures not captured by WER or CER alone.

2 Related Work

Multilingual and Indic ASR benchmarks. Large multilingual resources such as FLEURS (Conneau et al., 2023), ML-SUPERB (Shi et al., 2025), and Common Voice (Ardila et al., 2020) have broadened ASR evaluation beyond traditionally high-resource languages. Within the Indian context, IndicSUPERB/Kathbath (Javed et al., 2023), IndicVoices (Javed et al., 2024a), Vistaar and IndicWhisper (Bhogale et al., 2023), LAHAJA (Javed et al., 2024b), Vaani (Pulikodan et al., 2026), the Microsoft Indic speech corpus (Srivastava et al., 2020), and OpenSLR⁴ have expanded speech-resource coverage across several Indic languages. However, most existing benchmarks still focus on scheduled or comparatively well-resourced languages. Recent efforts such as MunTTS (Gumma et al., 2024) have started to address speech technology for Mundari, but both Bhili and Mundari continue to lack evaluation-ready ASR benchmarks with fixed splits, reproducible baselines, and diagnostics tailored to language-documentation settings.

Cross-lingual transfer and auxiliary supervision. Cross-lingual transfer is widely used for low-resource ASR. Multilingual models such as XLS-R (Babu et al., 2022), Whisper (Radford et al., 2023), and SeamlessM4T (Communication et al., 2023) provide strong initialization with limited target-language data, but may bias outputs toward dominant or better-resourced pretraining languages (Li et al., 2024). Low-resource machine translation exhibits analogous drift toward higher-resource related languages (Xu et al., 2020); in our ASR setting, this appears as Hindi-normalization,

⁴<https://www.openslr.org/>

Table 1: Coverage of Bhili/Mundari in major multilingual and Indic ASR resources. #L: languages covered; **Eval**: paired transcripts with reusable evaluation splits. [†]Unaligned audio only; no transcripts or baselines.

Resource	#L	Bhi	Mun	Eval
FLEURS	102	✗	✗	✓
IndicVoices	22	✗	✗	✓
IndicSUPERB	12	✗	✗	✓
Vistaar	12	✗	✗	✓
Vaani	54	3h [†]	✗	✗
AdiDhwani (ours)	2	56.5h	11.7h	✓

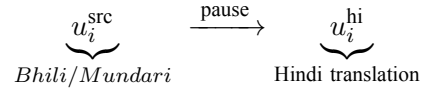
where community-attested Bhili forms may be replaced by Hindi-like cognates (Dhasmana et al., 2026). Synthetic speech and external language models are also widely used in low-resource ASR (Moslem, 2024; Thai et al., 2019; DeRenzi et al., 2025; Heafield, 2011; Liu et al., 2024), although gains depend on the match between auxiliary and target data (Bhatt and Bolonyai, 2011). We therefore evaluate donor-language initialization, synthetic supervision, and KenLM-assisted decoding as controlled conditions.

Evaluation beyond WER. WER and CER remain essential for standard ASR comparison, but they do not always capture errors that matter for language documentation (Kumar et al., 2026). Where the target language is in sustained contact with a dominant language, as Bhili and Mundari both are with Hindi (Appendix B), ASR systems may replace target-language forms with dominant-language cognates (Meyer et al., 2020; Koenecke et al., 2020); in morphologically rich languages, they may fragment agglutinated words or omit suffixes carrying grammatical meaning (K et al., 2025). Such errors may incur only a small edit-distance penalty while reducing the value of the transcript as a linguistic record. To make these failure modes directly measurable, we report four diagnostics (Section 4): dominant-language substitution in the hypothesis (HNB) or at the aligned reference position (HNBA), fragmentation of agglutinated forms (MFR), and deletion of suffixal material (SDR). A related but distinct line of work on endangered and Indigenous language technology (Le Ferrand et al., 2024) emphasizes consent, provenance, community benefit, and access governance (Bird, 2020; Mager et al., 2018; Nekoto et al., 2020; Duncan et al., 2024); these principles shape the collection and adjudication protocol in Section 3 rather than the diagnostics themselves.

3 Data Collection & Audio Preprocessing

Unlike a generic speech corpus, AdiDhwani preserves both target-language speech needed for ASR and an aligned Hindi translation tier used for annotation, auditability, and linguistic analysis.

Bilingual Elicitation. We collect speech using a bilingual prompt-and-translate protocol: each speaker first produces a target-language utterance and then provides its Hindi translation in the same recording, after a long pause.



The Hindi translation gives annotators an aligned semantic reference, while the target-first ordering discourages code-mixing in the segment used for ASR training. The translation tier is not used for ASR training or evaluation, but is retained as linguistic metadata for annotation, auditability, and future lexical analysis. Full field-recording and segmentation details are in Appendix C.

Segmentation & Filtering. Let $\mathbf{x}$ denote the raw continuous recording of a single elicitation item, comprising the target-language utterance, the intervening pause, and the Hindi translation. Although the pause was part of the recording instruction, it was not a fixed-duration acoustic marker: because sessions took place in community field settings rather than studios, realized pause length, speaker loudness, microphone distance, device gain, and background noise all varied across speakers and sessions. Splitting at a fixed offset would therefore truncate target-language speech after short pauses and retain Hindi translation after long ones. We instead detect silence intervals with energy-based VAD and take the longest as the target–Hindi boundary:

$$s^* = \arg \max_{s \in S(\mathbf{x})} \text{dur}(s), \quad (1)$$

where $S(\mathbf{x})$ is the set of silence intervals returned by the VAD. The audio before s^* is retained as the target-language segment. Segmentation and filtering serve distinct purposes: segmentation isolates the target-language segment used for ASR training and evaluation, whereas filtering is a subsequent quality-control step applying duration and SNR thresholds, boundary padding, and peak normalization. Manual verification of a 10% sample

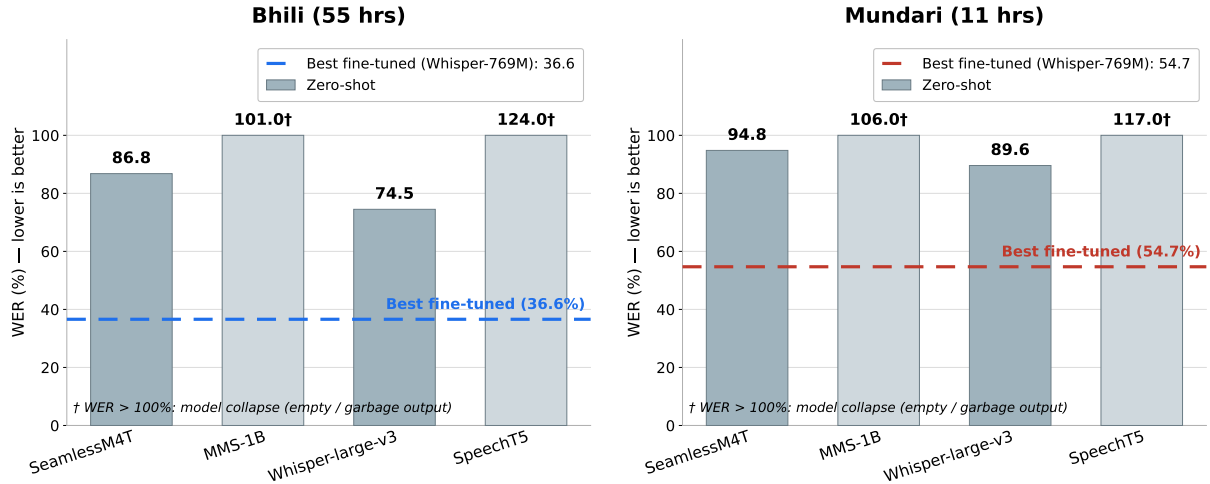


Figure 1: WER (%) on Bhili (55h) and Mundari (11h) before any fine-tuning, with best fine-tuned WER as dashed reference. Whisper-large-v3 achieves 74.5% and 89.6% WER respectively; MMS-1B and SpeechT5 hallucinated, scores reflect bad recognition.

Table 2: Corpus statistics for training and held-out test splits after segmentation and quality filtering.

Lang.	Split	Utt.	Dur. (h)	Mean (s)
Bhili	Train	19,000	55.0	10.5
	Test	558	1.5	10.2
Mundari	Train	7,500	11.0	5.3
	Test	500	0.7	5.1

from each session yielded a boundary error rate below 3%, indicating that automatic segmentation was reliable for constructing the target-language speech corpus. Algorithm 1 in Appendix C states the full procedure.

Annotation and Quality Control. Transcription follows a maker-checker-superchecker workflow involving community speakers and expert linguistic review. This is necessary because both languages present preservation-specific ambiguities: Bhili annotation is affected by Hindi-cognate normalization, while Mundari annotation requires consistent treatment of agglutinated forms and suffixal material. Superchecker intervention was required for 12.4% of Bhili utterances and 18.7% of Mundari utterances. Rather than treating these only as transcription errors, we view them as evidence of the sociolinguistic and morphological pressures that ASR systems must also handle. Annotation guidelines are detailed in Appendix D.

Auxiliary Synthetic Speech. No production-quality TTS system was available for Bhili and Mundari. We therefore generated auxiliary synthetic audio with Hindi Google TTS at a reduced speaking rate of 0.8, producing 12 h for Bhili and 3 h for Mundari. Synthetic audio is used only for training and never appears in development or test sets. We analyze its intelligibility, artifacts, and downstream effect in Appendices G, H and J.

4 Preservation-Oriented Evaluation

WER and CER are necessary for ASR comparison, but they treat all errors as linguistically equivalent. This is insufficient for preservation-oriented ASR, where a system may reduce edit distance while producing transcripts that are less faithful to the target language (Thomason and Kaufman, 2023; Myers-Scotton, 2002). We therefore report WER and CER for recognition accuracy, their ratio ρ as a compact summary of how the two relate, and four preservation diagnostics proposed in this work-HNB, HNBA, MFR, and SDR that target documentation-relevant failures directly. Construction details for diagnostics are in Appendix E.

WER, CER, and the CER/WER Ratio. For each language $\ell \in \{\text{Bhili}, \text{Mundari}\}$, WER measures word-level disagreement and CER measures character-level disagreement between the reference y_i^ℓ and the hypothesis $\hat{y}_i^\ell$ of model f ; i indexes the utterance. Alongside these, and following work that advocates reporting character- and word-level error jointly for multilingual and mor-

phonologically complex ASR (Morris et al., 2004; K et al., 2025), we report their ratio

$$\rho^\ell(f) = \frac{\text{CER}^\ell(f)}{\text{WER}^\ell(f)} \quad (2)$$

We treat ρ as a derived summary of two quantities we already report, not as one of the preservation diagnostics proposed in this work; unlike HNB, HNBA, MFR, and SDR (below), it introduces no new measurement and carries no fixed threshold. Its value is comparative: a lower ρ indicates that word errors tend to involve small character differences—word-boundary mistakes, lexical-selection errors, or orthographic variants, whereas a higher ρ indicates that word errors are more often accompanied by substantial character-level disagreement. ρ helps indicate whether residual errors are more likely to respond to language-model rescoring or to require stronger acoustic modelling. Across our post-LM systems Bhili stays lower (0.31–0.40) and Mundari higher (0.47–0.48); by midpoint, this is roughly one-third more character-level disagreement per word error for Mundari, consistent with its higher MFR and SDR (Section 6) and with the differing Bhili/Mundari responses to KenLM rescoring in Table 4.

Hindi-Normalization Bias. Let $\mathcal{B}_i^\ell$ be the monitored set for reference utterance i : a curated subset of target-language reference tokens flagged as at risk of Hindi-cognate substitution, rather than the full transcript vocabulary. Each token $t \in \mathcal{B}_i^\ell$ is one for which an approved Hindi alternative could plausibly surface in ASR output; these tokens are extracted from the bilingual target–Hindi tier by pointwise mutual information and confirmed by community reviewers (Appendix E). Let $\text{HindiAlt}_\ell(t)$ denote the approved Hindi-cognate alternatives for t . Define

$$h_\ell(t, \hat{y}_i^\ell) = \mathbf{1}[\exists z \in \text{HindiAlt}_\ell(t) : z \in \hat{y}_i^\ell]$$

For model f , Hindi-Normalization Bias is

$$\text{HNB}^\ell(f) = \frac{\sum_i \sum_{t \in \mathcal{B}_i^\ell} h_\ell(t, \hat{y}_i^\ell)}{\sum_i |\mathcal{B}_i^\ell|} \quad (3)$$

Because HNB may count incidental Hindi alternatives, we also report an alignment-constrained variant. For the hypothesis token $\hat{t}$ aligned to reference token t , define

$$h_{a,\ell}(t, \hat{t}) = \mathbf{1} \left[\begin{array}{l} t \in \mathcal{B}_i^\ell \wedge \\ \hat{t} \in \text{HindiAlt}_\ell(t) \end{array} \right]$$

Then

$$\text{HNBA}^\ell(f) = \frac{\sum_i \sum_{(t, \hat{t}) \in \text{align}(\hat{y}_i^\ell, \hat{y}_i^\ell)} h_{a,\ell}(t, \hat{t})}{\sum_i |\mathcal{B}_i^\ell|} \quad (4)$$

Thus, $\text{HNBA}^\ell \leq \text{HNB}^\ell$. These diagnostics are orthogonal to WER: a model may lower edit distance by producing a phonologically close Hindi cognate while increasing dominant-language normalization.

Morphological Fragmentation and Suffix Deletion. For both languages, we measure whether ASR hypotheses preserve agglutinated morphology. Let $\mathcal{A}_i^\ell$ be the set of agglutinated reference tokens in utterance i for language ℓ , identified using a community-validated suffix inventory. For an agglutinated reference token r , let $\hat{t}_r$ denote the hypothesis token aligned to r under the same WER edit-distance traceback used for HNBA, with $\hat{t}_r = \varepsilon$ when r is deleted.

For an agglutinated token r , define

$$M_\ell(r, \hat{y}_i^\ell) = \left[\begin{array}{l} \text{stem}_\ell(r) \in \hat{y}_i^\ell \wedge \\ \exists s \in \text{suffixes}_\ell(r) : s \in \hat{y}_i^\ell \wedge \\ \hat{t}_r \neq r \end{array} \right]$$

Then

$$m_\ell(r, \hat{y}_i^\ell) = \mathbf{1}[M_\ell(r, \hat{y}_i^\ell)]$$

The condition $\hat{t}_r \neq r$ excludes the case where the model reproduces r as a single correct token, which would otherwise be miscounted as fragmentation whenever the stem and a suffix both appear. Morphological Fragmentation Rate is

$$\text{MFR}^\ell(f) = \frac{\sum_i \sum_{r \in \mathcal{A}_i^\ell} m_\ell(r, \hat{y}_i^\ell)}{\sum_i |\mathcal{A}_i^\ell|} \quad (5)$$

Fragmentation captures cases where the stem and suffix are recognized but not preserved as a single agglutinated word.

Define the stem-preservation condition

$$S_\ell(r, \hat{y}_i^\ell) = \left[\begin{array}{l} \text{stem}_\ell(r) \in \hat{y}_i^\ell \vee \\ \hat{t}_r \text{ starts with } \text{stem}_\ell(r) \end{array} \right],$$

and the suffix-deletion condition

$$D_\ell(r, \hat{y}_i^\ell) = \left[\begin{array}{l} \nexists s \in \text{suffixes}_\ell(r) : \\ s \in \hat{y}_i^\ell \cup \{\hat{t}_r\} \end{array} \right]$$

Then

$$d_\ell(r, \hat{y}_i^\ell) = \mathbf{1}[S_\ell(r, \hat{y}_i^\ell) \wedge D_\ell(r, \hat{y}_i^\ell)]$$

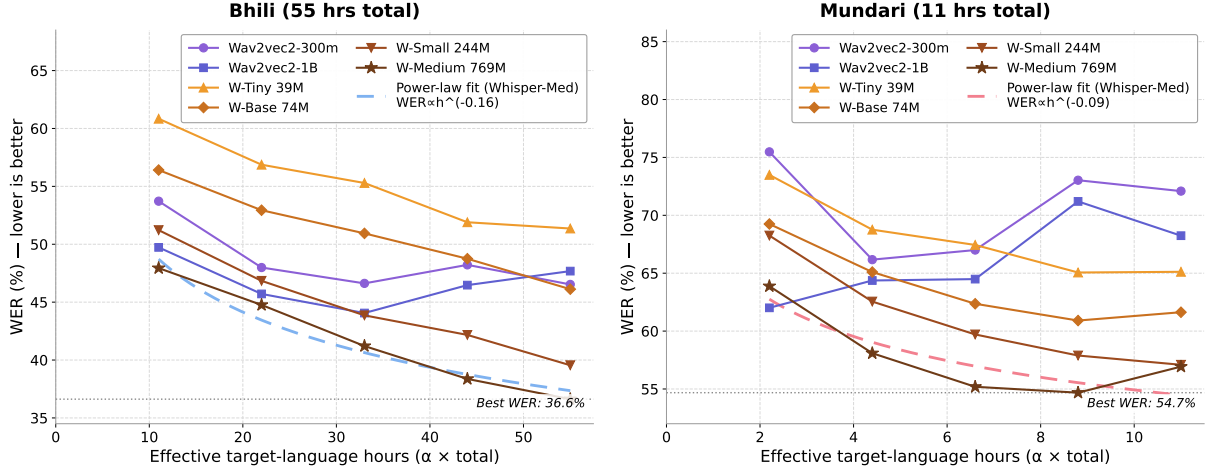


Figure 2: Target-language data scaling for Bhili and Mundari. WER is plotted against effective target-language hours, computed as α times the full training set size. The dashed curve is a power-law fit to Whisper-Medium; Bhili benefits from more data, while Mundari peaks before full-data training.

The suffix deletion rate is

$$\text{SDR}^\ell(f) = \frac{\sum_i \sum_{r \in \mathcal{A}_i^\ell} d_\ell(r, \hat{y}_i^\ell)}{\sum_i |\mathcal{A}_i^\ell|}. \quad (6)$$

SDR captures the more damaging case where the lexical stem survives but grammatical suffix material is lost. Together, MFR and SDR expose documentation-relevant errors that may be hidden by WER and CER, detailed in Appendix E.

5 Experimental Setup

Evaluation Protocol. We evaluate all systems on the fixed held-out test sets using a common transcript-preprocessing and scoring pipeline, with deterministic Unicode and orthographic normalization and consistent word-boundary conventions applied to both references and hypotheses, and dataset splits as shown in Table 2. WER is the primary ASR metric, with CER reported as a secondary character-level metric. We also report the CER/WER ratio and the preservation-oriented diagnostics defined in Section 4. Hyperparameters, and condition-wise results are provided in Appendix K and H along with human transcriber baselines in Appendix D.6.

Model Families. We evaluate models in both zero-shot and supervised settings. For zero-shot evaluation we cover the major publicly available multilingual speech paradigms—encoder-decoder ASR, CTC-style recognition, and unified speech/text models—using Whisper-large-v3 (Radford et al., 2023), SeamlessM4T-v2 (Communication et al., 2023), MMS-1B (Pratap et al., 2024),

and SpeechT5 (Ao et al., 2022). For supervised adaptation we fine-tune two complementary families: XLS-R 300M/1B (Babu et al., 2022) with CTC decoding, which supports the donor, acoustic-similarity, and KenLM analyses, and Whisper Tiny/Base/Small/Medium, which supports scaling comparisons across model size and target-data fraction. Since neither Bhili nor Mundari has an official Whisper language token, we use the Hindi token for all Whisper fine-tuning and decoding as mentioned in Appendix K.

Adaptation Regimes. We compare five adaptation regimes. First, we evaluate zero-shot transfer from multilingual ASR models. Second, we perform target-language fine-tuning, varying the target-language data fraction $\alpha \in \{0.2, 0.4, 0.6, 0.8, 1.0\}$, the proportion of the available target training set used, so that $\alpha=1.0$ is the full split. This corresponds to 11–55 effective target-language hours for Bhili and 2.2–11 hours for Mundari, and drives the scaling analysis in Appendix H.2. Third, we evaluate donor-language initialization: donor checkpoints are first trained on a candidate donor language and then adapted to the full target-language training split. Fourth, we test synthetic supervision through synthetic-only and synthetic-to-real training paths. We also evaluate the best XLS-R models with 5-gram language modeling. Full results are in Appendix H.

Donor Selection & Language-Model Decoding. We first shortlist donor languages using geographic and linguistic proximity, then rank

them using layer-wise acoustic similarity computed from XLS-R hidden states. This compares donor and target representations across acoustic (L1–6), phonetic (L7–12), and linguistic (L13–24) layers, giving a script-independent estimate of transfer compatibility. A candidate donor must also be *instantiable* as a controlled transfer condition: either sufficient transcribed speech to adapt XLS-R, or an established checkpoint for that language. This requirement, rather than genealogical distance, is why the Rajasthani varieties are absent from our candidate set despite belonging to the same Western Indo-Aryan continuum as Bhili; varieties such as Marwari and Mewari are themselves under-resourced by the criteria that motivate this benchmark. Appendix F.4 reports a zero-shot representational probe for Marwari together with the limitation that no fine-tunable donor checkpoint exists for these varieties. Donor initialization is evaluated primarily with XLS-R 300M CTC (Table 3); we additionally report Whisper-Medium in Appendix H.6 as an encoder–decoder architectural-sensitivity comparison. For CTC systems, we also train a 5-gram KenLM using only target-language training transcripts; development or test text are excluded from LM training. During decoding, CTC and LM scores are combined by shallow fusion. We do not apply KenLM to Whisper because its autoregressive decoder already provides an implicit language model. Full donor analysis is in Appendix F & H.

Table 3: Donor-transfer WER (%) using XLS-R 300M. Lower is better; the best donor for each language is shown in **bold**.

Language	Donor	WER (%)
Bhili	Hindi	45.33
	Marathi	57.91
Mundari	Hindi	64.13
	Santali	77.81

6 Results and Analysis

We organize the results into two complementary analyses. We first compare recognition performance across five adaptation regimes and identify the best-performing configuration for each language. We then apply the preservation-oriented diagnostics in Section 4 to examine whether gains in WER/CER also preserve documentation-relevant linguistic forms.

Zero-Shot Transfer. Whisper-large-v3 and SeamlessM4T produce valid Devanagari hypotheses, but remain far from usable accuracy (Figure 1). Whisper-large-v3 obtains 74.48% WER on Bhili and 89.62% WER on Mundari. MMS-1B and SpeechT5 often collapse into hallucinated or unstable outputs, showing that large multilingual pretraining does not guarantee graceful degradation on unseen under-documented languages. We therefore treat zero-shot performance as a stress test rather than a viable documentation system.

Target-Language Supervision. Fine-tuning sharply reduces WER for both languages, but the best use of target data depends on the resource setting (Figure 2). For Bhili, which has 55 hours of real speech, Whisper-Medium improves with more target-language data and reaches its best staged result at $\alpha=1.0$ with 36.62% WER. For Mundari, which has only 11 hours of real speech, Whisper-Medium performs best at $\alpha=0.8$ with 54.68% WER, while full target-only training degrades. Thus, target-language supervision is essential, but the optimal mixing ratio is not fixed: full target training helps in the 55-hour setting but can overfit in the 11-hour setting. We fit a power-law curve $WER = a \cdot h^{-b}$ to the Whisper-Medium results to summarize this scaling trend.

Effect of Synthetic Supervision. Synthetic-only training is substantially worse than staged real-data fine-tuning for both languages, indicating that TTS speech cannot replace real community speech. However, synthetic pretraining has language-dependent effects after real-data adaptation and LM decoding (Figure 3). For Bhili, where 55 hours of real speech are available, synthetic pretraining hurts the final LM path. For Mundari, where only 11 hours of real speech are available, the 3-hour synthetic stage becomes useful after KenLM rescoring. Synthetic speech is therefore most useful under severe data scarcity, and only when evaluated as part of a complete training-and-decoding pipeline.

Donor transfer depends on acoustic compatibility. Layer-wise donor analysis selects Hindi and Marathi as plausible donors for Bhili, and Hindi and Santali as plausible donors for Mundari as shown in Table 3, while rejecting Gujarati and Bengali despite geographic proximity (Appendix F). In downstream CTC transfer, Hindi is the only consistently helpful donor, including for Mundari even

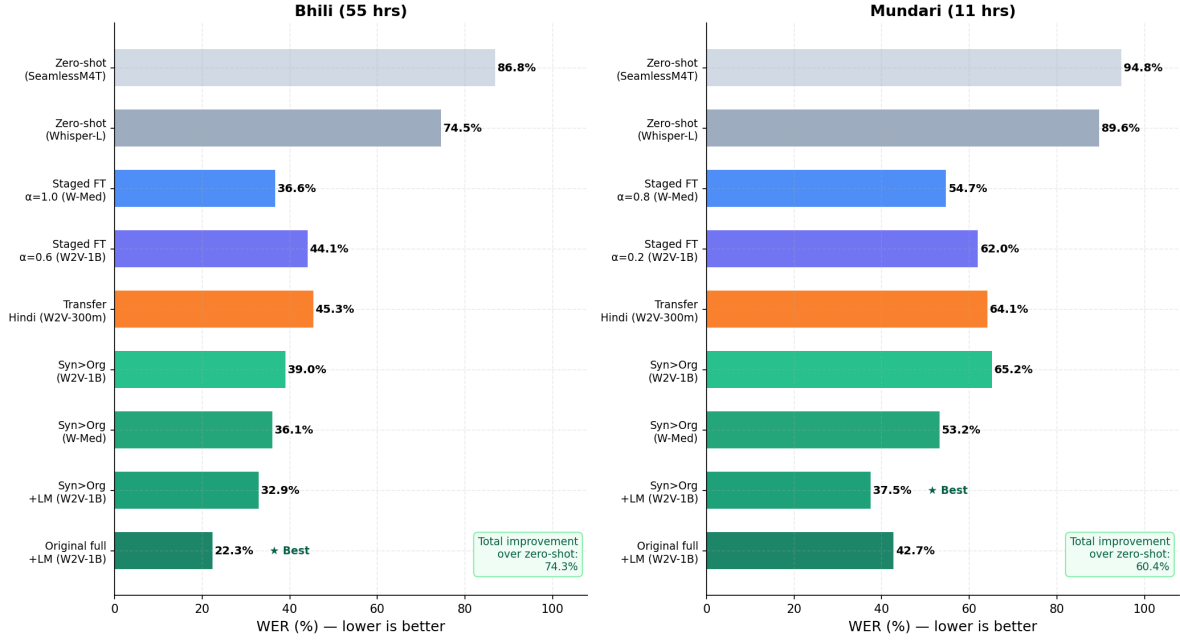


Figure 3: Best WER per training configuration for Bhili and Mundari, ordered weakest to strongest; stars mark the best system per language. Real-data fine-tuning with KenLM gives 22.3% for Bhili; Syn→Org+LM gives 37.5% for Mundari.

Table 4: Effect of KenLM shallow-fusion rescoring on real-only and synthetic-to-real training for Bhili and Mundari. The largest WER reductions are 21.74 pp for Bhili (Original+LM) and 27.66 pp for Mundari (Syn→Org+LM), which also achieves the best final Mundari WER (37.52%). Syn→Org: synthetic pre-training followed by real-data fine-tuning; Original: real-data only; +LM: KenLM shallow-fusion rescoring.

Language	Path	WER pre-LM	WER post-LM	Δ WER	CER post-LM	CER/WER
Bhili	Syn→Org+LM	39.00	32.91	−6.09	10.35	0.31
Bhili	Original+LM	44.06	22.32	−21.74	8.83	0.40
Mundari	Syn→Org+LM	65.18	37.52	−27.66	17.91	0.48
Mundari	Original+LM	62.01	42.74	−19.27	20.17	0.47

though Santali is genealogically closer. This suggests that acoustic representation similarity is a better predictor of ASR transfer than language-family membership alone.

Language-Specific Pipelines. We apply KenLM (Heafield, 2011; Liu et al., 2024) rescoring to the strongest XLS-R 1B CTC systems under two paths: real-data-only training and synthetic-to-real training. Table 4 shows that KenLM improves WER in all configurations, but the best post-LM path differs by resource condition. For Bhili, real-data-only training followed by KenLM is best, reaching 22.32% WER and 8.83% CER; it outperforms the Syn→Org+LM path by 10.59 WER points. With 55 hours of real speech, synthetic pre-training provides no additional benefit, suggesting that the acoustic mismatch from synthetic data may limit the gains from LM rescoring. For Mundari, the pattern

reverses. Although Syn→Org is worse than real-only training before LM, it becomes the best system after rescoring, reaching 37.52% WER and 17.91% CER. Under the 11-hour constraint, synthetic pre-training appears to produce hypotheses that are easier for the LM to rerank, even when the acoustic model alone is weaker. Thus, LM rescoring is not merely a uniform post-processing step; its effectiveness depends on the training path and available real-data resources. Detailed analysis is provided in Appendix H.

WER/CER Error Profiles. The CER/WER ratio separates word-level errors from deeper character-level acoustic confusion (Table 17). For Bhili, the lower ratios (0.31–0.40) indicate that many remaining errors involve lexical choice, orthographic variation, or word-boundary decisions that can be partly corrected by LM rescoring. For Mundari, both paths have higher ratios (0.47–

0.48), showing that many word errors also involve substantial character-level disagreement. KenLM therefore reduces WER, but it cannot fully address errors caused by limited acoustic evidence, especially for Mundari.

Preservation Fidelity Beyond WER. The preservation diagnostics in Table 5 show that strong ASR performance under WER does not necessarily imply faithful language documentation. For Bhili, the best system reaches 22.32% WER, while HNB remains 18.70% and alignment-constrained HNBA remains 17.09%. The small gap between HNB and HNBA indicates that many Hindi-normalized forms are direct substitutions at the aligned reference position, rather than incidental Hindi words elsewhere in the hypothesis. Thus, even a relatively accurate transcript may silently shift community-attested Bhili forms toward dominant Hindi variants. For Mundari, the best system still has 1.92% MFR and 4.47% SDR, showing that agglutinated forms are sometimes fragmented or reduced to stems without their suffixal material. Since such suffixes can encode grammatical information, these errors reduce the transcript’s value as a linguistic record even when the lexical stem is recognized. Human transcriber baselines are substantially stronger, with 11.24% WER for Bhili and 14.31% for Mundari, showing that these gaps are not inherent limitations of the languages. The best system by WER is therefore not necessarily the one that best preserves the language.

Table 5: Preservation-oriented diagnostics for the best ASR system per language and human transcriber baselines. Lower is better for all metrics; values are percentages.

Lang.	WER	HNB	HNBA	MFR	SDR
bhb (Human)	11.24	8.38	7.89	0.11	1.20
bhb	22.32	18.70	17.09	0.62	3.40
unr (Human)	14.31	10.17	9.18	0.32	2.17
unr	37.52	21.24	17.41	1.92	4.47

Metric orthogonality. A key observation is that the preservation diagnostics are not redundant with WER. Although Bhili has a 15.20 pp lower WER than Mundari, its alignment-constrained Hindi-normalisation bias is only 0.32 pp lower (17.09% vs. 17.41%). This means that better recognition accuracy does not necessarily remove dominant-language substitution at the token positions that matter for documentation. Conversely, Mundari

shows higher MFR and SDR, indicating a different failure mode: morphology is more often split or deleted even when the model recovers part of the lexical form. Human baselines reduce all four preservation errors, showing that these diagnostics capture real system limitations rather than unavoidable ambiguity in the data.

Linguistic Error Diagnostics. Beyond aggregate WER and CER, we analyze character-, word-, and utterance-level errors to diagnose the best ASR systems more precisely. The analysis separates normalisation artefacts, such as ZWNJ omission and visarga–colon mismatch, from documentation-relevant errors, including nukta loss, vowel-length confusion, Hindi postposition over-generation, and suffix deletion. Table 5 further compares each best system with human baselines using preservation-oriented diagnostics for Hindi-normalisation bias, alignment-constrained bias, morphological fragmentation, and suffix deletion. These analyses help identify whether remaining errors are associated with Unicode normalisation, acoustic modelling, LM rescoring, or limited linguistic diversity in the training data. Detailed analyses are provided in Appendices F, H, I.3, & J.

7 Conclusion

We introduce AdiDhwani, a preservation-oriented ASR benchmark for Bhili and Mundari, built through community-driven data collection and validation, with speaker-disjoint splits and reproducible baselines. Target-language supervision and language-model rescoring substantially improve recognition, although the best training strategy differs by resource condition: real-only training performs best for Bhili, while synthetic-to-real training with LM rescoring performs best for Mundari. Crucially, lower edit distance does not always correspond to better linguistic preservation: strong ASR systems still exhibit Hindi-normalisation bias in Bhili and suffix deletion in Mundari. We therefore complement WER and CER with HNB, HNBA, MFR, SDR, and the CER/WER ratio to expose preservation-relevant errors obscured by aggregate recognition metrics. These findings motivate evaluating ASR for under-documented languages not only by recognition accuracy, but also by how faithfully systems preserve linguistically meaningful forms.

Limitations

AdiDhwani is an initial benchmark for Bhili and Mundari, and several limitations remain. Although the corpus covers multiple districts and speaker groups, it does not capture the full sociolinguistic diversity of either language. Our donor analysis is also restricted to varieties with instantiable ASR conditions, which excludes the Rajasthani varieties most closely related to Bhili (Appendix F.4). Most recordings are read speech, with limited coverage of conversational and spontaneous interaction. Both languages are released primarily in Devanagari despite broader multi-script usage, particularly for Mundari. In addition, our experiments focus on fixed test sets and aggregate metrics, which limits fine-grained analysis across dialects, demographics, and downstream documentation settings. Future releases should expand speaker diversity, speech domains, script coverage, and community-led evaluation.

Ethics Statement

All participants were informed about the goals of the project, the recording process, and the intended research and documentation uses of the collected speech prior to participation. Instructions and consent procedures were provided in locally understandable languages, and informed verbal and written consent was obtained from all speakers. Participants were compensated in accordance with prevailing local daily wages to recognize their time, linguistic expertise, and contribution to the corpus.

The benchmark was developed in collaboration with community linguists, educators, and local language practitioners, who contributed to prompt selection and the handling of culturally sensitive content. Transcription quality followed a maker-checker-superchecker workflow with expert adjudication and correction logging to improve consistency and accountability. We frame AdiDhwani as a documentation-oriented research resource intended to support future work in language preservation, education, and community-centered speech technology.

Acknowledgments

This work was carried out under the AdiVaani initiative of the Ministry of Tribal Affairs, Government of India, supporting the development of language technologies and digital resources for tribal and endangered Indian languages. We gratefully

acknowledge the Ministry’s support to IIT Delhi, which enabled the research, language-resource development, community engagement, and computational infrastructure underlying this work. We also thank IIT Delhi HPC for providing the computational resources used in our experiments.

We thank Ms. Ranjana Chopra, Shri Anant Prakash Pandey, Ms. Deepali R. Masirkar, and Shri Ganesh Nagarajan (Ministry of Tribal Affairs); Prof. Vivek Kumar and Prof. Parag Singla (IIT Delhi); and Prof. Radhika Mamidi (IIIT Hyderabad) for their guidance and institutional support. We are grateful to the Tribal Research Institutes (TRIs), local institutions, and community partners in Madhya Pradesh and Jharkhand for facilitating engagement with the Bhili and Mundari language communities.

For the Bhili resource in Jhabua, Madhya Pradesh, we gratefully acknowledge Sarika Dhoulpuria, Arun Khawse, Bajrang Singh Nalwaya, Prakash Amliyar, Kanju Damor, Jawsingh Dhak, Madan Singad, Rajesh Bawariya, Vinod Guniya, Gul Singh Gohil, Mangla Garwal, Indira Gundia, Kamlesh Bilwal, Dhansingh Chauhan, Raju Katara, and Lokesh Singad. For the Mundari resource in Jharkhand, we gratefully acknowledge Vikas Dubey, Alok Shikari Kachhap, Ruby Kumari, Raju Toppo, Soma Singh Munda, Etwamunda, Ajit Munda, Birsingh Munda, Bisnu Mundri, Kuldeep Dang, Basanti Kumari Munda, Amar Topno, Birsa Purty, Jagdish Soy, Baiju Munda, Lekhan Prakash Munda, Lebyan Mundu, Serofina Hemrom, Saroj Soymurmn, Sunita Tidu, Rachna Horo, Saloni Mundu, Sudha Manisha Nag, Virjman Munda, Shamu Nag, Monika Kumari, Tulshi Hari Munda, Ramsay Munda, Divya Topno, Priyanka Aind, Babli Kumari, Leel Mohan Munda, Kande Munda, Roshan Purty, Anupam Purty, Deet Orea, Ram Dodrai, Jaimany Barauid, Ajay Kerketta, Vickey Toppo, and Anima Kachap. We thank these language experts and community contributors for the linguistic knowledge, care, and sustained effort that supported the development and validation of the resources used in this study.

We also thank Amit Kumar, Mohammad Fazal (IIT Delhi); Anindita Mondal and Nikhilesh (IIIT Hyderabad), Mithun Chandravanshi, Praveen, Nicolette, and Anuj Tandon (Ministry of Tribal Affairs) for their support in coordinating the broader project activities. We further acknowledge the researchers, students, developers, and technical

teams who contributed to the recording, annotation, validation, data-management, and computational infrastructure supporting this work.

Most importantly, we thank the Bhili and Mundari speakers and their communities for contributing their speech, time, and linguistic knowledge. Their participation is fundamental to AdiDhwani as a community-grounded resource for endangered-language speech technology and documentation.

References

- Benjamin Akera, Evelyn Nafula, Patrick Walukagga, Gilbert Yiga, John Quinn, and Ernest Mwebaze. 2025. How much speech data is necessary for asr in african languages? an evaluation of data scaling in kinyarwanda and kikuyu. *arXiv preprint arXiv:2510.07221*.
- Gregory DS Anderson. 2015. *The Munda Languages*. Routledge.
- Junyi Ao, Rui Wang, Long Zhou, Chengyi Wang, Shuo Ren, Yu Wu, Shujie Liu, Tom Ko, Qing Li, Yu Zhang, Zhihua Wei, Yao Qian, Jinyu Li, and Furu Wei. 2022. [SpeechT5: Unified-modal encoder-decoder pre-training for spoken language processing](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5723–5738, Dublin, Ireland. Association for Computational Linguistics.
- Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis Tyers, and Gregor Weber. 2020. Common voice: A massively-multilingual speech corpus. In *Proceedings of the twelfth language resources and evaluation conference*, pages 4218–4222.
- Arun Babu, Changhan Wang, Andros Tjandra, Kushal Lakhotia, Qiantong Xu, Naman Goyal, Kritika Singh, Patrick von Platen, Yatharth Saraf, Juan Pino, Alexei Baevski, Alexis Conneau, and Michael Auli. 2022. [XLS-R: Self-supervised Cross-lingual Speech Representation Learning at Scale](#). In *Interspeech 2022*, pages 2278–2282.
- Laurent Besacier, Etienne Barnard, Alexey Karpov, and Tanja Schultz. 2014. Automatic speech recognition for under-resourced languages: A survey. *Speech communication*, 56:85–100.
- Rakesh M Bhatt and Agnes Bolonyai. 2011. Code-switching and the optimal grammar of bilingual language use. *Bilingualism: Language and cognition*, 14(4):522–546.
- Kaushal Bhogale, Sai Sundaresan, Abhigyan Raman, Tahir Javed, Mitesh M Khapra, and Pratyush Kumar. 2023. Vistaar: Diverse benchmarks and training sets for indian language asr. In *Proc. Interspeech 2023*, pages 4384–4388.
- Steven Bird. 2020. Decolonising speech and language technology. In *Proceedings of the 28th international conference on computational linguistics*, pages 3504–3519.
- Seamless Communication, Loïc Barrault, Yu-An Chung, Mariano Cora Meglioli, David Dale, Ning Dong, Paul-Ambroise Duquenne, Hady Elsahar, Hongyu Gong, Kevin Heffernan, John Hoffman, Christopher Klaiber, Pengwei Li, Daniel Licht, Jean Maillard, Alice Rakotoarison, Kaushik Ram Sadagopan, Guillaume Wenzek, Ethan Ye, and 49 others. 2023. [Seamlessm4t: Massively multilingual & multimodal machine translation](#). *Preprint*, arXiv:2308.11596.
- Alexis Conneau, Min Ma, Simran Khanuja, Yu Zhang, Vera Axelrod, Siddharth Dalmia, Jason Riesa, Clara Rivera, and Ankur Bapna. 2023. Fleurs: Few-shot learning evaluation of universal representations of speech. In *2022 IEEE Spoken Language Technology Workshop (SLT)*, pages 798–805. IEEE.
- Christopher Cox. 2011. Corpus linguistics and language documentation: challenges for collaboration. *Language & Computers*, 73(1).
- Brian DeRenzi, Anna Dixon, Mohamed Aymane Farhi, and Christian Resch. 2025. Synthetic voice data for automatic speech recognition in african languages. In *Proceedings of the First Workshop on Advancing NLP for Low-Resource Languages*, pages 152–186.
- Akriti Dhasmana, Aaroohi Srivastava, and David Chiang. 2026. [Dialect matters: Cross-lingual ASR transfer for low-resource Indic language varieties](#). In *Proceedings of the 13th Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 145–156, Rabat, Morocco. Association for Computational Linguistics.
- Suzanne Duncan, Gianna Leoni, Lee Steven, Keoni Mahelona, and Peter-Lucas Jones. 2024. Fit for our purpose, not yours: Benchmark for a low-resource, indigenous language. *Advances in Neural Information Processing Systems*, 37:31543–31561.
- Asma Etman and AA Louis Beex. 2015. Language and dialect identification: A survey. In *2015 SAI intelligent systems conference (IntelliSys)*, pages 220–231. IEEE.
- Varun Gumma, Rishav Hada, Aditya Yadavalli, Pamir Gogoi, Ishani Mondal, Vivek Seshadri, and Kalika Bali. 2024. Muntts: A text-to-speech system for mundari. In *Proceedings of the Seventh Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 76–82.
- Kenneth Heafield. 2011. Kenlm: Faster and smaller language model queries. In *Proceedings of the sixth workshop on statistical machine translation*, pages 187–197.

- Tahir Javed, Kaushal Bhogale, Abhigyan Raman, Pratyush Kumar, Anoop Kunchukuttan, and Mitesh M Khapra. 2023. Indicsuperb: A speech processing universal performance benchmark for indian languages. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pages 12942–12950.
- Tahir Javed, Janki Nawale, Eldho George, Sakshi Joshi, Kaushal Bhogale, Deovrat Mehendale, Ishvinder Sethi, Aparna Ananthanarayanan, Hafsa Faquih, Pratiti Palit, Sneha Ravishankar, Saranya Sukumaran, Tripura Panchagnula, Sunjay Murali, Kunal Gandhi, Ambujavalli R, Manickam M, C Vijayanthi, Krishnan Karunganni, and 2 others. 2024a. [IndicVoices: Towards building an inclusive multi-lingual speech dataset for Indian languages](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 10740–10782, Bangkok, Thailand. Association for Computational Linguistics.
- Tahir Javed, Janki Nawale, Sakshi Joshi, Eldho George, Kaushal Bhogale, Deovrat Mehendale, and Mitesh M Khapra. 2024b. Lahaja: A robust multi-accent benchmark for evaluating hindi asr systems. *arXiv preprint arXiv:2408.11440*.
- Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the nlp world. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 6282–6293.
- 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.
- Suyoun Kim, Abhinav Arora, Duc Le, Ching-Feng Yeh, Christian Fuegen, Ozlem Kalinli, and Michael L Seltzer. 2021. Semantic distance: A new metric for asr performance analysis towards spoken language understanding. *arXiv preprint arXiv:2104.02138*.
- Allison Koenecke, Andrew Nam, Emily Lake, Joe Nudell, Minnie Quartey, Zion Mengesha, Connor Toups, John R Rickford, Dan Jurafsky, and Sharad Goel. 2020. Racial disparities in automated speech recognition. *Proceedings of the national academy of sciences*, 117(14):7684–7689.
- Rishabh Kumar, Amrith Krishna, Ganesh Ramakrishnan, and Preethi Jyothi. 2026. [Post-ASR correction in Hindi: Comparing language models and large language models in low-resource scenarios](#). In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 636–645, Rabat, Morocco. Association for Computational Linguistics.
- Éric Le Ferrand, Raina Heaton, and Emily Prud’hommeaux. 2024. Enenlhet as a case-study to investigate asr model generalizability for language documentation. In *Proceedings of the 4th Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP 2024)*, pages 132–137.
- Zhaolin Li, Monika Rind-Pawłowski, and Jan Niehues. 2024. Speech recognition corpus of the khinalug language for documenting endangered languages. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 15171–15180.
- Constantine Lignos and Mitch Marcus. 2013. Toward web-scale analysis of codeswitching. In *87th Annual Meeting of the Linguistic Society of America*, volume 90.
- Zoey Liu, Nitin Venkateswaran, Éric Le Ferrand, and Emily Prud’hommeaux. 2024. How important is a language model for low-resource asr? In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 206–213.
- Manuel Mager, Ximena Gutierrez-Vasques, Gerardo Sierra, and Ivan Meza-Ruiz. 2018. Challenges of language technologies for the indigenous languages of the americas. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 55–69.
- Josh Meyer, Lindy Rauchenstein, Joshua D Eisenberg, and Nicholas Howell. 2020. Artie bias corpus: An open dataset for detecting demographic bias in speech applications. In *Proceedings of the twelfth language resources and evaluation conference*, pages 6462–6468.
- Andrew Cameron Morris, Viktoria Maier, and Phil Green. 2004. From wer and ril to mer and wil: improved evaluation measures for connected speech recognition. In *Proc. Interspeech 2004*, pages 2765–2768.
- Yasmin Moslem. 2024. Leveraging synthetic audio data for end-to-end low-resource speech translation. In *Proceedings of the 21st International Conference on Spoken Language Translation (IWSLT 2024)*, pages 265–273.
- Carol Myers-Scotton. 2002. *Contact linguistics: Bilingual encounters and grammatical outcomes*. OUP Oxford.
- Alvin Nahabwe, Sulaiman Kagumire, Denis Musinguzi, Bruno Beijuka, Jonah Mubuke Kyagaba, Peter Nabende, Andrew Katumba, and Joyce Nakatumba-Nabende. 2025. Benchmarking automatic speech recognition models for african languages. *arXiv preprint arXiv:2512.10968*.
- Wilhelmina Nekoto, Vukosi Marivate, Tshinondiwa Matsila, Timi Fasubaa, Taiwo Fagbohunge,

- Solomon Oluwole Akinola, Shamsuddeen Muhammad, Salomon Kabongo Kabenamualu, Salomey Osei, Freshia Sackey, Rubungo Andre Niyongabo, Ricky Macharm, Perez Ogayo, Orevaghene Ahia, Musie Meressa Berhe, Mofetoluwa Adeyemi, Masabata Mokgesi-Seling, Lawrence Okegbemi, Laura Martinus, and 28 others. 2020. [Participatory research for low-resourced machine translation: A case study in African languages](#). In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2144–2160, Online. Association for Computational Linguistics.
- Timothy Osborne, Michael Putnam, and Thomas Groß. 2012. Catenae: Introducing a novel unit of syntactic analysis. *Syntax*, 15(4):354–396.
- Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: an asr corpus based on public domain audio books. In *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5206–5210. IEEE.
- 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. 2024. [Scaling speech technology to 1,000+ languages](#). *Journal of Machine Learning Research*, 25(97):1–52.
- Sujith Pulikodan, Abhayjeet Singh, Agneeth Basu, Nihar Desai, Pavan Kumar J, Pranav D Bhat, Raghu Dharmaraju, Ritika Gupta, Sathvik Udupa, Saurabh Kumar, Sumit Sharma, Visruth Sanka, Dinesh Tewari, Harsh Dhand, Amrita Kamat, Sukhwinder Singh, Shikhar Vashishth, Partha Talukdar, Raj Acharya, and Prasanta Kumar Ghosh. 2026. [Vaani: Capturing the language landscape for an inclusive digital india](#). *Preprint*, arXiv:2603.28714.
- Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518. PMLR.
- Jiatong Shi, Dan Berrebbi, William Chen, Ho-Lam Chung, En-Pei Hu, Wei Ping Huang, Xuankai Chang, Shang-Wen Li, Abdelrahman Mohamed, Hung yi Lee, and Shinji Watanabe. 2025. [Ml-superb: Multilingual speech universal performance benchmark](#). *Preprint*, arXiv:2305.10615.
- Nimisha Srivastava, Rudrabha Mukhopadhyay, Pranjwal K R, and C V Jawahar. 2020. [IndicSpeech: Text-to-speech corpus for Indian languages](#). In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6417–6422, Marseille, France. European Language Resources Association.
- Bao Thai, Robert Jimerson, Dominic Arcoraci, Emily Prud’hommeaux, and Raymond Ptucha. 2019. Synthetic data augmentation for improving low-resource asr. In *2019 IEEE Western New York image and signal processing workshop (WNYISPW)*, pages 1–9. IEEE.
- Sarah Grey Thomason and Terrence Kaufman. 2023. *Language contact, creolization, and genetic linguistics*. Univ of California Press.
- Changhan Wang, Morgane Riviere, Ann Lee, Anne Wu, Chaitanya Talnikar, Daniel Haziza, Mary Williamson, Juan Pino, and Emmanuel Dupoux. 2021. Voxpopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 993–1003.
- Jin Xu, Xu Tan, Yi Ren, Tao Qin, Jian Li, Sheng Zhao, and Tie-Yan Liu. 2020. Lrspeech: Extremely low-resource speech synthesis and recognition. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2802–2812.

Appendix Overview

A Linguistic Background of Bhili and Mundari

B Field Recording Protocol

C Annotation Guidelines and Adjudication

D Preservation Metric Construction and Calibration

E Donor Language Selection

F Synthetic Speech Analysis and HNB Risk

G Expanded Result & Data Analysis

H Linguistic Metric Results and Analysis

I Linguistic Analysis of ASR Systems

J Training & Compute Infrastructure

K Cultural Representation and Dataset Governance

L Generalizing Beyond the Bhili–Mundari Benchmark

A Data and Code Release

The AdiDhwani datasets and associated resources are publicly released to support reproducibility and responsible reuse.

Hosting and access. Code and evaluation resources are available at <https://github.com/misniitdelhi-ai/AdiDhwani>. The Bhili and Mundari speech recordings, transcripts, and fixed data splits are hosted at <https://huggingface.co/datasets/misniitd/AdiDhwani>, where access instructions and conditions of use are documented.

License and governance. The speech recordings and transcripts are released under CC BY-NC-SA 4.0⁵. Access and reuse follow the consent and community-governance conditions documented with the dataset release.

Released artifacts. The release comprises: (i) train and test splits with audio and corresponding transcripts; (ii) WER/CER scoring code, including the micro-averaging convention and deterministic Unicode normalization; (iii) implementations of HNB, HNBA, MFR, and SDR, together with their associated lexical and morphological resources; and (iv) training and decoding configurations, including the KenLM order and shallow-fusion weight.

B Linguistic Background of Bhili and Mundari

This section provides linguistic context for the preservation-oriented evaluation used in AdiDhwani. For both languages, the central risks are dominant-language normalization toward Hindi and the loss of agglutinated morphology, especially suffixal material.

B.1 Bhili Linguistic Profile

Bhili (bhb) is a Western Indo-Aryan language continuum spoken across tribal regions of Gujarat, Rajasthan, Madhya Pradesh, and Maharashtra. The benchmark covers 56.6 hours of Bhili speech from 12 speakers across two districts. Because Bhili is in sustained contact with Hindi and other regional Indo-Aryan languages, it shares substantial phonological and lexical overlap with them. This overlap is useful for cross-lingual ASR transfer, but it also creates a preservation risk: a multilingual ASR system may replace a community-attested Bhili form with a more frequent Hindi cognate while incurring only a small edit-distance penalty.

⁵<https://creativecommons.org/licenses/by-nc-sa/4.0/>

B.2 Mundari Linguistic Profile

Mundari (unr) belongs to the Munda branch of the Austroasiatic family and is spoken across the Chota Nagpur Plateau region of eastern India. Jharkhand is its principal linguistic region, with substantial speaker communities also found in Odisha, West Bengal, and Bihar, and smaller populations in Chhattisgarh (Anderson, 2015). The benchmark contains 11.7 hours of Mundari speech from seven speakers across two districts. Unlike Bhili, whose main ASR challenge is Indo-Aryan contact and Hindi-like lexical substitution, Mundari presents a strongly morphological challenge. Its agglutinative structure allows multiple grammatical categories to be encoded within a single surface word, so ASR systems may recognise a stem while fragmenting the word or silently deleting suffixal material.

B.3 Contact, Script, and Evaluation Implications

Hindi contact affects both languages, but in different ways. In Bhili, it creates a direct risk of lexical normalization. In Mundari, it affects transcription practice and bilingual elicitation, but the larger structural challenge is preserving morphology. Both languages are represented in the benchmark using a consistent Devanagari-based transcription layer, a benchmark design decision rather than a complete representation of community writing practices. For Bhili, a system can reduce WER by substituting a Hindi-like cognate instead of the community form; for Mundari, a system can recognise a stem while deleting suffixes that carry grammatical meaning. This is why WER alone is insufficient for endangered-language ASR evaluation, motivating the diagnostics introduced in Section 4.

C Field Recording Protocol

The Bhili and Mundari corpora were collected through community-centered fieldwork in native-speaker communities. Community coordinators and local language practitioners supported speaker recruitment, consent procedures, and session organization. Speakers were informed about the purpose of the recordings, the intended research and documentation uses of the data, and the role of the Hindi translation tier. Written and verbal consent were obtained in locally accessible languages before recording.

Recordings were captured at 16 kHz, 16-bit

mono using directional microphones. Sessions were conducted in semi-controlled field conditions such as schools, community halls, and outdoor gathering spaces. Because the data were collected in natural community environments rather than studios, the recording protocol included real-time quality monitoring and post-hoc quality filtering for noise, clipping, and segmentation failures.

Data collection used a custom Android-based recording platform designed for low-resource and intermittently connected field environments. The platform supports offline-first recording, local encrypted storage, delayed synchronization, and integrated quality flagging by field annotators. This allowed field teams to continue recording in locations with unreliable network access while preserving metadata and session-level quality notes.

C.1 Voice Activity Detection and Boundary Detection

We apply energy-based voice activity detection independently to each recording. Short-time energy is computed using 25 ms windows with a 10 ms hop. A frame is labelled silent if its energy falls below a file-adaptive threshold, defined as 40 dB below the 99th-percentile frame energy for that recording. Adjacent silent frames are merged into contiguous silence intervals. A file-adaptive threshold is used because field recordings vary in speaker loudness, microphone distance, device gain, and background noise as shown in algorithm 1. Because the realized pause is variable rather than fixed, and each file contains exactly one target–translation pair, segmentation reduces to identifying the acoustic boundary between the target-language utterance and the Hindi translation. Let $\mathcal{S}(\mathbf{x}) = \{s_1, \dots, s_k\}$ be the set of silence intervals detected in recording $\mathbf{x}$. We select the longest silence interval between two audio samples:

$$s^* = \arg \max_{s \in \mathcal{S}(\mathbf{x})} \text{dur}(s). \quad (7)$$

The audio before s^* is retained as the target-language utterance, while the audio after s^* is retained as Hindi translation metadata. In verified files, this maximum-pause rule was unambiguous in over 97% of cases.

C.2 Quality Filtering and Manual Verification

After segmentation, extracted target-language segments pass four filters: utterances shorter than

Algorithm 1 Silence-based segmentation of bilingual field recordings.

Require: Raw recording $\mathbf{x}$ containing one target–translation pair

Ensure: Target-language utterance u^{src} , or Reject

- 1: Apply energy-based VAD to $\mathbf{x}$
 - 2: Extract silence intervals $\mathcal{S}(\mathbf{x}) = \{s_1, \dots, s_k\}$
 - 3: Select $s^* = \arg \max_{s \in \mathcal{S}(\mathbf{x})} \text{dur}(s)$
 - 4: $u^{\text{src}} \leftarrow$ audio before s^*
 - 5: $u^{\text{hi}} \leftarrow$ audio after s^*
 - 6: Preserve u^{hi} as translation metadata
 - 7: Reject if $\text{dur}(u^{\text{src}}) \notin [0.5, 20]$ seconds
 - 8: Estimate SNR for u^{src}
 - 9: Reject if $\text{SNR}(u^{\text{src}}) < 10$ dB
 - 10: Add 50 ms silence padding at both boundaries
 - 11: Peak-normalize u^{src} to -3 dBFS
 - 12: **return** filtered u^{src}
-

0.5 s are removed; utterances longer than 20 s are removed; utterances with estimated SNR below 10 dB are discarded; and accepted utterances are padded with 50 ms of silence at both boundaries. All accepted utterances are peak-normalized to -3 dBFS to reduce variation caused by speakers, devices, microphone distance, and field sessions while preserving speech content.

A random 10% sample from each recording session was manually inspected by a fluent community member or trained reviewer. Reviewers checked that (i) the retained segment contained only target-language speech, (ii) no target-language speech was truncated at segment boundaries, and (iii) the Hindi translation tier was separated correctly. The estimated boundary error rate was below 3% across sessions. Corrected boundaries from this verification sample were not used to tune the automatic segmentation procedure, so the reported error rate provides a conservative estimate of segmentation quality on the unverified portion of the corpus.

D Annotation Guidelines and Adjudication

Target-language prompts were produced from Hindi source sentences drawn from public datasets and translated by experts with more than 20 years of experience in the respective languages. The translated text was validated by senior supercheckers before being used for audio recording. Final transcripts were then produced and adjudicated through a community-centered annotation workflow.

D.1 Annotator Profiles and Training

Annotation was carried out by trained local annotators, community language practitioners, and senior reviewers. For Bhili, twelve trained annotators and two senior supercheckers participated in the final transcription process. For Mundari, seven trained annotators and one senior superchecker participated. Annotators were selected based on language competence, familiarity with local speech varieties, and ability to follow shared transcription conventions. Before contributing to the benchmark, annotators completed a one-week training programme followed by a 100-utterance calibration phase. Training covered the recording protocol, annotation interface, orthographic conventions, common sources of disagreement, and procedures for flagging uncertain cases.

D.2 Prompt Design

Prompts were developed to reflect community-relevant language use rather than isolated generic sentences. The prompt inventory was intentionally weighted toward governance and welfare access, since many practical translation needs in these communities involve public schemes, rights, local administration, and access to services. Education was kept as a second major domain because school communication, learning materials, and employment-related language are frequent use cases for community-facing translation. The remaining prompts cover agriculture, healthcare, oral history, and ecological knowledge, ensuring that the benchmark also captures everyday and documentation-oriented language use.

Community linguists reviewed the prompts for lexical authenticity, domain relevance, and cultural appropriateness. Prompts with unnatural phrasing, ambiguous local terminology, or domain-inappropriate wording were revised or removed before translation. This process ensured that the benchmark evaluates socially grounded language use rather than only surface-level sentence translation. Tables 6 and 7 summarize the domain distribution of accepted prompts after review and filtering.

For endangered and under-documented languages, these domains are not peripheral to speech and translation technology evaluation. They are central to the documentation settings that the benchmark is intended to support, including access to public information, educational participa-

Table 6: Bhili prompt distribution by documentation domain. Counts cover all 19,558 elicitation prompts and show the intended balance across public services, education, agriculture, health, oral history, and ecological knowledge.

Domain	Topic	Count
Governance/Welfare	Schemes, rights, services	9,779
Education	School, learning, employment	3,912
Agriculture	Crops, irrigation	1,467
Healthcare	Symptoms, nutrition	1,467
Oral history	Testimony, festivals	1,466
Ecological knowledge	Environment, local knowledge	1,467
Total	–	19,558

Table 7: Mundari prompt distribution by documentation domain. Counts cover all 8,000 elicitation prompts, with emphasis on governance and welfare topics.

Domain	Topic	Count
Governance/Welfare	Schemes, rights, services	4,000
Education	School, learning, employment	1,600
Agriculture	Crops, irrigation	600
Healthcare	Symptoms, nutrition	600
Oral history	Testimony, festivals	600
Ecological knowledge	Environment, local knowledge	600
Total	–	8,000

tion, oral histories, ecological knowledge, local administration, agricultural work, and community-facing communication.

D.3 General Transcription Principles

We followed the same transcription principles for both languages. Annotators were asked to preserve the speaker’s lexical choice whenever it was intelligible, rather than replacing it with a more formal or dominant-language equivalent. They were also asked not to normalize speech toward Hindi or another familiar regional form. Repetitions, hesitations, and self-repairs were retained when they were acoustically clear and belonged to the target utterance. When a segment of audio was unclear, annotators marked it for review instead of guessing. We also encouraged consistent word boundaries across recurring forms so that similar spoken forms were represented consistently in the transcripts.

These principles are important because replacing Bhili forms with Hindi cognates, or omitting weakly articulated Mundari suffixes, would make the transcript less faithful to the target language.

D.4 Language-Specific Guidelines

For Bhili, annotators were instructed to keep community-attested forms when speakers produced them, even if a Hindi-like form would be more familiar to an outside transcriber. This was necessary because Bhili is in sustained contact with Hindi and other regional languages, and many forms are phonologically close to Hindi cognates.

For Mundari, annotators were instructed to preserve suffixes and agglutinated forms when they functioned as a single phonological and grammatical unit. The guidelines discouraged splitting an agglutinated form into multiple orthographic tokens unless the speaker clearly produced the material as separate words. Weakly articulated suffixes were not omitted when they were recoverable from the audio and consistent with the utterance. Uncertain cases were escalated to the checker or superchecker rather than silently simplified.

D.5 Maker–Checker–Superchecker Workflow

Transcription followed a three-stage workflow.

Maker. The maker recorded the target-language audio or produced the first transcript. The maker could consult the Hindi translation tier for semantic orientation, but the transcript itself had to reflect the target-language speech.

Checker. The checker independently reviewed the maker transcript against both the target-language audio and the Hindi translation tier. The checker flagged uncertain words, possible Hindi-normalized forms, unclear suffixes, code-switching, repetitions, and cases where the audio did not match the intended prompt.

Superchecker. The superchecker resolved disagreements, adjudicated ambiguous forms, and maintained consistency with community-validated spelling and morpheme-boundary conventions.

D.6 Human Transcriber Baseline

To establish an achievable-performance reference for the preservation diagnostics, we collected independent human transcriptions of the evaluation data. Five transcribers for Bhili and two for Mundari re-transcribed the *complete* held-out test set for each language: 558 utterances (1.5 h) for Bhili and 500 utterances (0.7 h) for Mundari, as reported in Table 2. Human and system outputs are therefore evaluated on exactly the same utterances.

Transcribers were fluent community speakers who worked from audio alone, without access to the Hindi translation tier, elicitation prompt, or system outputs, and followed the transcription principles in Appendices D.3 and D.4.

Human transcripts were scored against the superchecker-adjudicated reference using the same normalization and scoring pipeline as the ASR systems, including the identical monitored-token set $\mathcal{B}^\ell$ and suffix inventory $\mathcal{A}^\ell$. Because the reference itself was produced through the maker–checker–superchecker workflow described in Appendix D.5, the human baseline measures agreement with the adjudicated transcription convention rather than an absolute performance ceiling. We therefore interpret it as evidence that the observed preservation gaps are not intrinsic to the languages, rather than as a theoretical lower bound on achievable error.

E Preservation Metric Construction and Calibration

This section describes how the preservation-oriented diagnostics introduced in Section 4 are constructed and calibrated. The formal metric definitions are given in Equations 3–6; here we provide the lexicon, suffix-inventory, alignment, and interpretation details needed to reproduce the measurements. These diagnostics are not substitutes for WER and CER. Instead, they identify documentation-relevant errors that edit distance alone may obscure.

E.1 Hindi-Cognate Lexicon Construction

For each utterance i , the monitored set $\mathcal{B}_i^\ell$ contains community-reviewed target-language tokens for which a Hindi alternative could plausibly appear in ASR output. These tokens are selected from the bilingual target–Hindi elicitation tier and then reviewed by language consultants.

For each Hindi token h and target-language token t , we first estimate their association using pointwise mutual information:

$$\text{PMI}(h, t) = \log \frac{p(h, t)}{p(h)p(t)}. \quad (8)$$

A candidate Hindi alternative is retained when

$$c(h, t) \geq 2 \quad \text{and} \quad \text{PMI}(h, t) \geq 0.5.$$

Hindi function words are removed before review, since they frequently produce false positives in

short utterances. The remaining mappings are manually checked by community experts. During this review, we remove cases where the Hindi and target-language forms do not represent a documentation-relevant substitution. The resulting set is used as $\text{HindiAlt}_\ell(t)$ in Equations 3 and 4. This monitored set is distinct from the full transcript vocabulary, which contains all reference words, and from the suffix inventory used for MFR and SDR (Appendix E.3, Suffix Inventory Construction), which targets agglutinated morphology rather than Hindi-normalization.

E.2 Alignment-Constrained Hindi Normalization

The unconstrained HNB score counts whether a Hindi alternative appears anywhere in the hypothesis. This captures broad Hindi drift, but it can also count incidental Hindi words that are not direct substitutions for the monitored reference token. To separate these cases, HNBA uses the reference–hypothesis alignment produced by the WER edit-distance traceback.

In HNBA, a Hindi-normalization event is counted only when the hypothesis token aligned to a monitored reference token is one of that token’s approved Hindi alternatives. Insertions and unrelated Hindi words elsewhere in the hypothesis therefore affect HNB but not HNBA. By construction,

$$\text{HNBA}^\ell(f) \leq \text{HNB}^\ell(f),$$

because every alignment-constrained substitution is also an unconstrained Hindi-alternative occurrence, but not every unconstrained occurrence is aligned to the monitored reference token.

E.3 Suffix Inventory Construction

For MFR and SDR, agglutinated reference tokens are identified using community-validated suffix inventories. The Bhili inventory contains more than 30 suffixes. The Mundari inventory contains more than 25 suffixes, including endings, directional suffixes, and subject-focus markers.

A reference token r is treated as agglutinated only if a known suffix can be removed while leaving an attested stem of at least two characters. Cases are excluded when the stem boundary is ambiguous, when the remaining form is not an attested stem, or when the analysis is rejected by the superchecker. This conservative filtering reduces the chance that ordinary lexical variation is

mistaken for morphological fragmentation or suffix deletion.

E.4 Morphological Error Classification

For each agglutinated token $r \in \mathcal{A}_i^\ell$, we compare the reference token with the aligned hypothesis token and with the full hypothesis string.

MFR counts cases where the stem and suffixal material are both present in the hypothesis but are not preserved as the original agglutinated word. This typically corresponds to word-boundary errors, where the ASR system recognizes parts of the form but splits or rearranges the morphology.

SDR counts the more severe case where the lexical stem survives but the suffixal material is absent. Stem preservation is allowed either as a standalone hypothesis token or as the prefix of the aligned hypothesis token. Suffix deletion is counted only when no known suffix for r appears in the aligned token or elsewhere in the hypothesis. Thus, SDR focuses on cases where the transcript preserves lexical identity while losing grammatical information encoded by the suffix.

E.5 Calibration and Interpretation

A useful ASR system for language documentation should reduce WER and CER while also avoiding increases in HNB, HNBA, MFR, and SDR. HNB and HNBA are interpreted as dominant-language substitution rates over the monitored token set. A small HNB–HNBA gap indicates that most Hindi alternatives are direct aligned substitutions. A larger gap indicates more incidental Hindi co-occurrence in the hypothesis.

MFR and SDR measure different forms of morphological degradation. MFR captures fragmentation of agglutinated forms, while SDR captures loss of suffixal grammatical material. We treat SDR as the more damaging error type because the transcript may appear lexically plausible while omitting grammatical information important for documentation.

F Donor Language Selection

This appendix describes how donor languages were selected for cross-lingual ASR transfer. Geographic proximity and language-family membership are useful starting points, but they are not sufficient predictors of ASR transfer. A nearby language may still have poor acoustic compatibility with the target language, while a dominant contact language may provide stronger acoustic repre-

sentations because of larger training resources or stronger phonetic overlap.

F.1 Candidate Donor Languages

For Bhili, we considered Hindi, Marathi, and Gujarati. Hindi was included because of sustained contact and shared Indo-Aryan structure as shown in Table 9. Marathi was included because some Bhili-speaking regions are in contact with Marathi. Gujarati was included as a geographically adjacent contrastive donor. Rajasthani varieties were considered but could not be instantiated as donor conditions; see Appendix F.4.

For Mundari, we considered Hindi, Santali, and Bengali as shown in Table 10. Hindi was included because of contact, and the availability of stronger Hindi ASR resources. Santali was included because it is genealogically closest among the candidates, as both Santali and Mundari belong to the Munda branch of Austroasiatic. Bengali was included as a geographically relevant contrastive donor.

F.2 Layer-wise Acoustic Similarity Method

We rank candidate donors using layer-wise acoustic similarity computed from XLS-R 300M hidden states. For each language a separate model was trained and used to get the token representation. For each candidate donor, the same target-language audio is encoded by the donor-adapted XLS-R model. The resulting hidden representations are compared with the corresponding target-language representations using cosine similarity as shown in Table 8. Layers are grouped into four zones: CNN feature encoder, acoustic zone (L1–6), phonetic zone (L7–12), and linguistic zone (L13–24). The phonetic zone is the most useful signal for donor suitability in our experiments.

Table 8: Donor selection decisions based on XLS-R 300M representational similarity. Donors with phonetic-zone similarity above 0.6 are retained as transfer candidates; Gujarati (0.067) and Bengali (0.063) are rejected despite regional proximity, confirming that phonetic-layer similarity predicts transfer utility better than geography.

Target	Donor	Acoustic (L1–6)	Phonetic (L7–12)	Justification
Bhili	Hindi	0.838	0.751	Shared family, contact, Devanagari
Bhili	Marathi	0.833	0.673	Regional contact, retroflex/nasal inventory
Mundari	Santali	0.867	0.825	North Munda, shared phonation types
Mundari	Hindi	0.841	0.870	Contact language, shared transcription script
Bhili	Gujarati (rejected)	0.168	0.067	Low phonetic similarity
Mundari	Bengali (rejected)	0.157	0.063	Low phonetic similarity

F.3 Donor Similarity Results

Table 9: Layer-zone cosine similarity between Bhili audio and candidate donor languages using XLS-R 300M hidden states. Scores are averaged over the CNN feature encoder, acoustic layers (L1–6), phonetic layers (L7–12), and linguistic layers (L13–24). Higher similarity indicates stronger donor compatibility; Hindi is the closest donor across all zones.

Donor	CNN	Acoustic (L1–6)	Phonetic (L7–12)	Linguistic (L13–24)
Hindi	0.982	0.838	0.751	0.392
Marathi	0.911	0.833	0.673	0.259
Gujarati	0.262	0.168	0.067	0.014

For Bhili, Figures 6 and 7, Hindi has the highest similarity across all zones, including a phonetic-zone similarity of 0.751. Marathi is also plausible, with a strong phonetic-zone score of 0.673 and an acoustic-zone score close to Hindi. Gujarati, despite geographic adjacency, has very low acoustic and phonetic similarity and is treated as a rejected donor.

For Mundari, Figure 8, 9, both Santali and Hindi are plausible donors. Santali has the highest acoustic-zone similarity, which is expected because Santali and Mundari are both Munda languages. Hindi has the highest phonetic-zone and linguistic-zone similarity, likely reflecting contact effects, shared Devanagari transcription in the benchmark, and stronger Hindi acoustic representations. Bengali has low phonetic-zone similarity and negative linguistic-zone similarity, so it is rejected despite regional proximity.

Across both target languages, donors with phonetic-zone similarity above approximately 0.6 behave as plausible transfer candidates, while donors below 0.1 are weak candidates. We treat this as a practical heuristic for this benchmark, not as a universal threshold. Layer-wise similarity is used as a donor-selection signal, not as a guarantee of downstream improvement. Final donor choice should still be validated through target-language adaptation.

F.4 Rajasthani Varieties as Candidate Donors

Bhili sits within the Western Indo-Aryan continuum alongside the Rajasthani varieties, so relatedness alone would make Marwari, Mewari, Dhundhari, Hadothi, or Shekhawati plausible donors. We exclude them on grounds of experimental instantiability rather than linguistic distance, and report

Table 10: Layer-zone cosine similarity between Mundari audio and candidate donor languages using XLS-R 300M hidden states. Scores are averaged over the CNN feature encoder, acoustic layers (L1–6), phonetic layers (L7–12), and linguistic layers (L13–24). Higher scores indicate stronger donor compatibility; Hindi and Santali are close in lower layers, while Hindi remains strongest in the higher linguistic zone.

Donor	CNN	Acoustic (L1–6)	Phonetic (L7–12)	Linguistic (L13–24)
Hindi	0.985	0.841	0.870	0.432
Santali	0.982	0.867	0.825	0.362
Bengali	0.236	0.157	0.063	−0.009

here both the resource constraint and a partial probe of the one variety for which any evaluation is currently possible.

Instantiability constraint. The procedure in Appendix F adapts XLS-R to a candidate donor, encodes Bhili audio with the donor-adapted model, and compares the resulting representations against Bhili’s own. This requires either enough transcribed speech to fine-tune XLS-R, as for Hindi, Marathi, Gujarati, Santali, and Bengali, or an established open-source checkpoint for the variety in question; our Hindi donor adapter, for reference, was trained on 100 hours of transcribed speech. Neither condition currently holds for any Rajasthani variety. To our knowledge no transcribed corpus is distributed through OpenSLR, IndicVoices, Vistaar, or IndicSUPERB, and no established XLS-R checkpoint exists. Project Vaani has collected Marwari speech, but the publicly reported transcribed portion remains under approximately 10 hours, well short of what is needed to train a donor adapter under the same protocol. Treating “Rajasthani” as a single donor would further conflate dialect mixture, corpus provenance, and transcription convention with linguistic relatedness.

The constraint is itself informative. These varieties are extremely low-resource by exactly the criteria that motivate AdiDhwani: they are candidate *targets* for the methodology described here rather than usable donors, and could serve as donors only after the corpus-construction effort this paper undertakes for Bhili and Mundari.

Zero-shot representational probe. Marwari is the only Rajasthani variety for which we could obtain speech at all; Mewari and the remaining va-

rieties have no publicly available audio and cannot be probed. For Marwari, what remains feasible without a donor adapter is a probe within a fixed representation space. We encode Bhili, Marwari, and Gujarati audio through our fine-tuned Bhili model and measure layer-wise cosine similarity to Bhili’s own representations, using the layer zones defined in Appendix F. Because all three languages pass through a single Bhili-adapted encoder rather than through separate donor-adapted models, these values are not comparable with Tables 9 and 10; only the ordering among the three languages is interpretable. Cosine similarity between speech representations is dominated by a shared speech-like component that inflates all scores, so we additionally report a mean-centred variant in which the mean representation across the compared languages is subtracted at each layer before similarity is computed. Table 11 reports the outcome. In the phonetic zone (L7–12), the zone that proved most discriminative for donor selection (Appendix F), Marwari falls below Gujarati, the donor already rejected in Table 8 and mean-centring preserves the ordering while widening the gap. The margin is small, so we read the two measurements as agreeing in direction rather than as establishing a precise distance; the probe yields no evidence that Marwari is more acoustically compatible with Bhili than a candidate we reject on the same zone.

Table 11: Zero-shot representational probe: phonetic-zone (L7–12) similarity to Bhili audio for Marwari, relative to Gujarati, with all languages encoded by the fine-tuned Bhili model. Marwari falls below the already-rejected Gujarati donor under both the raw and the mean-centred computation. Because a Bhili-adapted rather than donor-adapted encoder is used, these values are not comparable with Tables 9 and 10.

Comparison	Raw	Mean-centred
Marwari vs. Gujarati (phonetic)	−0.94%	−2.3%

The probe measures representational proximity within a Bhili-adapted space and is not equivalent to the donor-adapted measurement used for the retained candidates, which cannot be performed for Marwari without a fine-tunable checkpoint. It therefore constrains the plausibility of these varieties as donors without settling the question: their transfer utility remains untested, and we leave it to future work once dialect-labelled speech resources become available.

G Synthetic Speech Analysis and HNB Risk

Synthetic speech is not treated as a substitute for real community speech. It is evaluated as auxiliary supervision for low-resource ASR adaptation.

G.1 Synthetic Data Generation and Intelligibility

At corpus construction time, no production-quality TTS system was available for Bhili, Mundari, or Santali. We therefore generated auxiliary speech using Google Hindi TTS with the voice hi-IN-Wavenet-A. The speaking rate was reduced to $0.8\times$ to improve intelligibility and reduce deletion errors in synthesized speech. For Bhili, we synthesized 12 hours of auxiliary speech from approximately 4,000 filtered sentences. For Mundari, we synthesized 3 hours from approximately 2,000 filtered sentences. Synthetic utterances are used only in training experiments and never appear in development or test sets.

We assessed the synthetic audio using mean opinion score (MOS) ratings collected from native speakers under controlled listening conditions. Six raters evaluated the Bhili set and three the Mundari set, with 250 samples rated per language on a five-point scale. Both sets were rated as intelligible: 4.86 ± 0.08 for Bhili and 3.92 ± 0.21 for Mundari. Raters noted, however, that the accent was mismatched in both languages, describing the output as a Hindi speaker reading Bhili or Mundari text. Intelligibility therefore does not imply accent fidelity, and we treat the synthetic audio as auxiliary supervision rather than as a substitute for community speech.

G.2 Hindi TTS and Normalization Risk

Using Hindi TTS introduces a preservation risk: it may encourage Hindi-like pronunciations or decoder priors. This risk is most important where many target forms are phonologically close to Hindi cognates. We therefore interpret synthetic supervision through both recognition and preservation metrics. A synthetic training path is not considered a pure gain unless it improves WER/CER without worsening preservation behavior. Morphological structure is a second risk: Hindi TTS does not reliably realize agglutinated suffixal material, so training on synthetic speech may under-represent exactly the forms that MFR and SDR are designed to measure.

Table 12: Zero-shot WER (%) on Bhili and Mundari before target-language adaptation. Lower is better. More than 100% WER for MMS-1B and SpeechT5 indicates degenerated, hallucinated or near-empty decoding outputs.

Model	Bhili	Mundari
Whisper-large-v3	74.48	89.62
SeamlessM4T-v2	86.84	94.77
MMS-1B	101	106
SpeechT5	124	117

H Expanded Results & Data Analysis

H.1 Zero-Shot Baseline Analysis

We first evaluate four pre-trained multilingual ASR systems in a zero-shot setting, without fine-tuning on the target languages. This establishes a baseline for how off-the-shelf systems behave on previously unseen tribal languages. Results are shown in Figure 1 and Table 12.

A key observation is that low WER does not always indicate successful recognition. The near-zero WER for MMS-1B (101%) and SpeechT5 (124%) should not be interpreted as strong performance. Instead, it reflects **model collapse**: the systems often output hallucinated strings, repeated tokens, or a single familiar word. This produces artificially low edit distance while failing to recognize the utterance. This is an important negative result. Although MMS-1B was trained on 1000+ languages, it fails on Bhili and Mundari because these languages are absent from its training data.

These results show that large-scale multilingual pretraining does not guarantee graceful degradation on unseen languages. SeamlessM4T and Whisper-large-v3 perform poorly (74–95% WER), but they still produce coherent output in related languages, making them plausible starting points for fine-tuning. By contrast, the collapse behavior of MMS-1B and SpeechT5 makes them unsuitable fine-tuning candidates for zero-resource tribal languages in this setting.

H.2 Scaling Law Analysis

Methodology: Because we do not have separate incremental training runs at different corpus sizes, we use the **alpha hyperparameter** as a proxy for effective target-language data. At $\alpha = 0.2$, target-language data contributes 20% of the training mix; at $\alpha = 1.0$, it contributes 100%. This corresponds to the following effective hours:

- **Bhili:** 11h, 22h, 33h, 44h, 55h (at $\alpha = 0.2, 0.4, 0.6, 0.8, 1.0$)
- **Mundari:** 2.2h, 4.4h, 6.6h, 8.8h, 11h

We fit a power-law curve $WER = a \cdot h^{-b}$ to the Whisper-Medium results, as shown in Figure 2, Table 13 & 14.

Bhili Scaling Behavior: For Bhili (55 hours total), Whisper-Medium follows a near-consistent downward trend, as reported in Figure 2. The curve does **not saturate** within 55 hours, suggesting that Bhili has not yet reached a data wall. The fitted power-law exponent $b \approx 0.09$ and 0.16 are consistent with findings in low-resource NLP (Akera et al., 2025). This suggests that collecting even 20–30 additional hours of Bhili speech would likely yield meaningful gains.

Mundari Scaling Behavior: For Mundari (11 hours), the trend is less stable, especially for wav2vec2. Whisper-Medium performs best at $\alpha = 0.8$ (54.68%) and then **degrades** at $\alpha = 1.0$ (56.93%). With only 11 hours of real data, training entirely on target-language data is harmful. The model benefits from the regularization introduced by donor-language mixing; more details are provided in Appendix F.

This points to a **language-resource-dependent optimal alpha**. For the higher-resource setting, Bhili (55h), $\alpha = 1.0$ is optimal. For the lower-resource setting, Mundari (11h), $\alpha = 0.8$ helps reduce overfitting.

Architectural Scaling Differences: Overall, the results point to a **language-resource-dependent optimal alpha**. For the higher-resource setting, Bhili (55h), $\alpha = 1.0$ is optimal. For the lower-resource setting, Mundari (11h), $\alpha = 0.8$ helps reduce overfitting. The scaling behavior also differs by architecture: Whisper models benefit more consistently from additional target-language data, whereas wav2vec2 models show **non-monotonic behavior**, especially on Mundari. Wav2vec2-1B is worse than Wav2vec2-300M on Mundari at most alpha values (62.01 vs. 66.17 at $\alpha = 0.2$, changing to 68.24 vs. 72.09 at $\alpha = 1.0$). Larger CTC models may overfit faster in scarce-data settings because they have more capacity to memorize the small training set. This architecture-dependent scaling behaviour mirrors findings for African low-resource ASR, where the relative ranking of pretrained models shifts

with the amount of available target-language data (Nahabwe et al., 2025).

H.3 Alpha Sensitivity Analysis

Figure 5 shows that the optimal value of α depends on both the language and the model family. For Bhili, the two wav2vec2 models perform best with a mixed training regime: Wav2vec2-1B reaches its best WER at $\alpha = 0.6$ (44.06%), and Wav2vec2-300M peaks at $\alpha = 1.0$ (46.53%). In contrast, Whisper models benefit from using all available target-language data. Whisper-Medium achieves its best result at $\alpha = 1.0$ (36.62%), as does Whisper-Small (39.55%). This indicates that, in the 55-hour Bhili setting, Whisper can use additional target-language supervision more effectively than the CTC-based wav2vec2 models.

For Mundari, the preferred mixing ratios shift toward lower values, which is consistent with the much smaller 11-hour training set. Wav2vec2-300M is best at $\alpha = 0.4$ (66.17%), while Wav2vec2-1B performs best at $\alpha = 0.2$ (62.01%). This unusually low optimum suggests that the larger CTC model is highly sensitive to overfitting when exposed to too little target-language data. Whisper models are more stable, with their best results occurring broadly around $\alpha = 0.8$. Overall, wav2vec2 requires more donor-language mixing for Mundari, whereas Whisper can tolerate a larger share of target-language data before overfitting.

Sensitivity Range: The WER range across alphas, computed as worst minus best, indicates how sensitive each model is to the mixing ratio. Results are summarized in Table 14. Whisper models are **more sensitive to alpha** than wav2vec2 models, as shown by their larger ranges. This means that choosing a poor alpha value has a higher cost for Whisper. A likely reason is that the Whisper decoder balances target-language generation against cross-lingual knowledge, and alpha directly changes this balance, as shown in Figure 12. These results provide a systematic alpha sensitivity study for staged fine-tuning of large multilingual ASR models on tribal languages, showing that the optimal value is not always 1.0.

H.4 Model Size vs. WER: Pareto Efficiency

Figure 4 compares model size against the best WER obtained by each model family. At comparable parameter counts, Whisper consis-

Table 13: WER (%) under staged target-language fine-tuning for Bhili and Mundari. The fraction α denotes the proportion of available target-language training data used in each stage: $\alpha=1.0$ corresponds to 55 hours for Bhili and 11 hours for Mundari. Lower WER is better; the best value for each model and language is shown in bold.

Arch.	Model	Bhili					Mundari				
		0.2	0.4	0.6	0.8	1.0	0.2	0.4	0.6	0.8	1.0
XLS-R	300M	53.72	47.99	46.62	48.22	46.53	75.48	66.17	67.00	73.03	72.09
	1B	49.73	45.71	44.06	46.47	47.68	62.01	64.37	64.49	71.20	68.24
Whisper	Tiny	60.84	56.87	55.29	51.90	51.36	73.49	68.76	67.44	65.06	65.12
	Base	56.41	52.94	50.94	48.75	46.12	69.25	65.10	62.36	60.91	61.63
	Small	51.22	46.84	43.86	42.17	39.55	68.25	62.55	59.71	57.89	57.09
	Medium	47.94	44.76	41.21	38.37	36.62	63.88	58.09	55.19	54.68	56.93

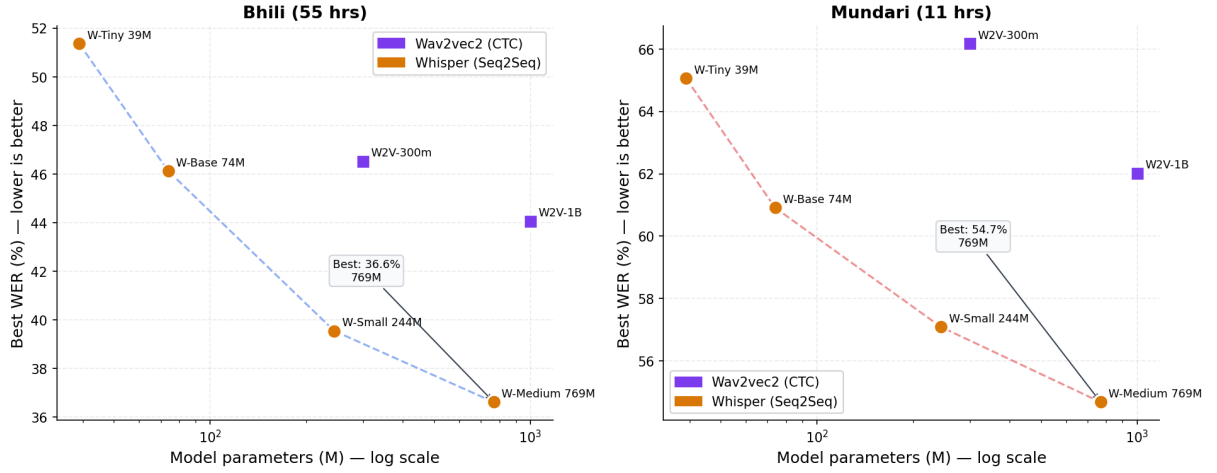


Figure 4: Model parameter count (log scale) versus best WER for Bhili (left) and Mundari (right); squares = XLS-R CTC, circles = Whisper seq2seq, dashed line = Pareto frontier. Whisper-Small (244M, 39.6%) is Pareto-efficient over XLS-R 1B (1B params, 44.1%), showing architecture matters more than scale.

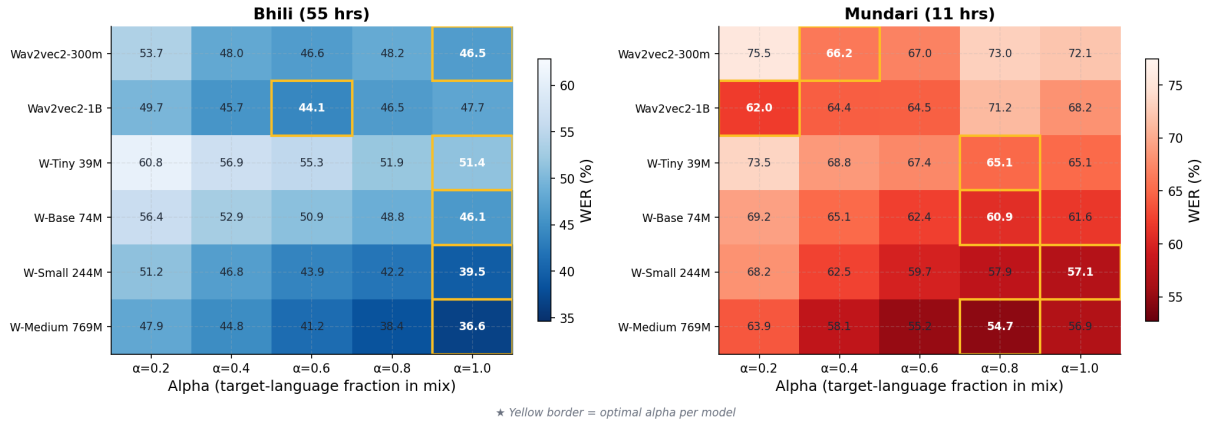


Figure 5: WER heatmap across data-mixing ratios $\alpha \in \{0.2, 0.4, 0.6, 0.8, 1.0\}$ for all models on Bhili (left) and Mundari (right); yellow borders mark the per-model optimal α , darker cells indicate lower WER. No single α is universally optimal: Whisper prefers $\alpha=1.0$ for Bhili but $\alpha=0.8$ for Mundari.

tently gives better WER than wav2vec2 for Bhili. Around the 300M scale, Wav2vec2-300M achieves 46.53% WER, while Whisper-Small reaches 39.55%. Around the 1B scale, Wav2vec2-1B reaches 44.06%, while Whisper-

Medium reaches 36.62%. For Mundari, the gap is smaller but the same pattern holds at the medium scale: Wav2vec2-1B reaches 62.01%, compared with 54.68% for Whisper-Medium.

The Pareto frontier also favors Whisper. For

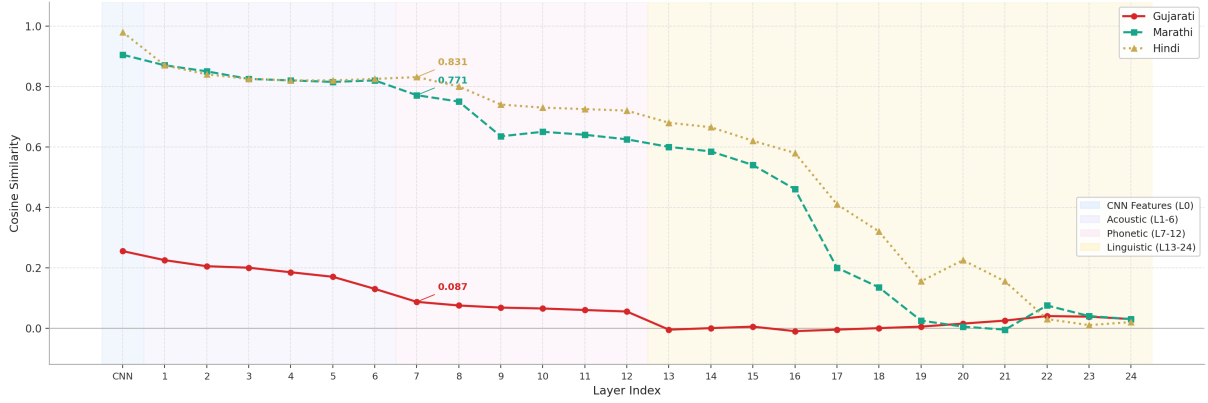


Figure 6: Layer-wise cosine similarity between Bhili audio and Hindi, Marathi, and Gujarati donor representations across all 25 XLS-R 300M layers. Hindi is the strongest donor (phonetic-zone similarity 0.751); Gujarati drops to 0.067 after the CNN encoder despite geographic adjacency and is rejected.

Table 14: WER range (pp) across $\alpha \in \{0.2, 0.4, 0.6, 0.8, 1.0\}$ per model, measuring sensitivity to the data-mixing ratio. Whisper models show larger ranges than XLS-R (up to 11.67 pp), meaning a poorly chosen α imposes a higher cost for seq2seq than CTC architectures.

Model	Bhili range	Mundari range
XLS-R 300M	7.19	9.31
XLS-R 1B	5.67	9.19
Whisper Tiny	9.48	8.43
Whisper Base	10.29	8.34
Whisper Small	11.67	11.16
Whisper Medium	11.32	9.20

Bhili, Whisper-Small (244M, 39.55%) is Pareto-efficient over Wav2vec2-1B (1B parameters, 44.06%), giving better WER with roughly $4\times$ fewer parameters. For Mundari, Whisper-Medium (769M, 54.68%) dominates the wav2vec2 variants. This suggests that, under the tested conditions, model architecture matters at least as much as parameter count: larger CTC models do not automatically translate into better low-resource ASR performance.

The scaling gains also show diminishing returns. For Bhili, moving from Whisper-Base to Whisper-Small improves WER from 46.12% to 39.55%, a 6.57 pp gain for 170M additional parameters. Moving from Whisper-Small to Whisper-Medium improves WER only from 39.55% to 36.62%, a 2.93 pp gain for 525M additional parameters. This pattern suggests that Bhili is already in a **data-bound regime**: adding more speech data would likely help Whisper-Medium more than increasing model size further. For Mundari, where only 11 hours of real speech are available, the data con-

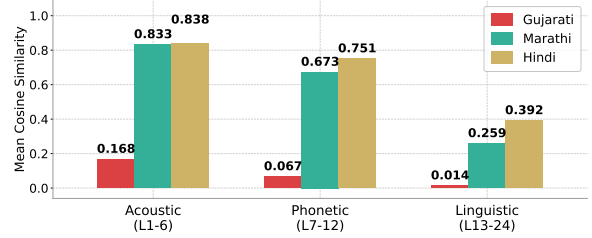


Figure 7: Zone-averaged cosine similarity between Bhili and candidate donors across acoustic (L1–6), phonetic (L7–12), and linguistic (L13–24) XLS-R layers. Hindi leads all three zones; the phonetic zone is the most discriminative for donor selection, separating Hindi and Marathi from rejected Gujarati.

straint is even stronger.

H.5 Synthetic Data Analysis

Figure 10 shows that synthetic speech is not a substitute for staged fine-tuning on real data. Synthetic-only training is much worse for both languages: 74.08% WER for Bhili compared with 36.62% under staged fine-tuning, and 79.5% WER for Mundari compared with 56.93%. These gaps indicate substantial domain mismatch in the TTS-generated speech. Although synthetic audio is clean and may help the model associate acoustic patterns with linguistic tokens, it does not capture the coarticulation, prosody, and speaker variability of real community speech.

Using synthetic data as a pre-training step before real-data fine-tuning (Syn→Org) gives a more nuanced result. For Bhili with XLS-R 1B, the staged baseline improves from 44.06% to 39.00%, a 5.06 pp gain. This suggests that 55 hours of real speech are sufficient for the model to correct

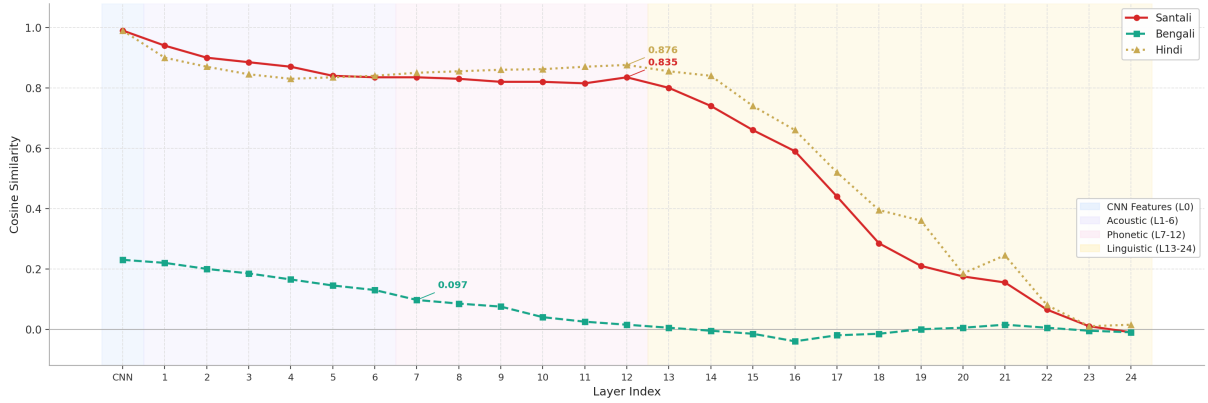


Figure 8: Layer-wise cosine similarity between Mundari audio and Hindi, Santali, and Bengali donor representations across all 25 XLS-R 300M layers. Santali and Hindi are both viable donors in lower layers; Bengali collapses below 0.1 after the CNN encoder and is rejected despite regional proximity.

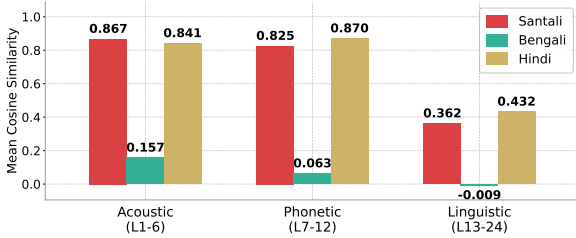


Figure 9: Zone-averaged cosine similarity between Mundari and candidate donors. Santali leads the acoustic zone (0.867) via Munda-family proximity, but Hindi leads the phonetic zone (0.870) and is the stronger downstream transfer donor, confirming that acoustic similarity predicts transfer better than genealogy.

much of the TTS-induced bias during fine-tuning. For Mundari, however, the same strategy degrades performance from 62.01% to 65.18%. With only 11 hours of real speech, the model does not have enough target-language evidence to overcome the acoustic bias introduced by synthetic pre-training.

The exchange rate between synthetic and real data is therefore unfavorable in this setting. For Bhili, the 5.06 pp gain from Syn→Org requires 12 hours of synthetic speech; based on the scaling curve, roughly 11 additional hours of real speech would provide a similar improvement. This makes synthetic data approximately $1.09\times$ less efficient than real data in this pipeline. Overall, synthetic speech is useful only when enough real target-language data is available to correct TTS artifacts during the final fine-tuning stage. Below that threshold, as in the 11-hour Mundari setting, synthetic pre-training can hurt unless it is followed by LM integration (see Section H.7).

H.6 Transfer Learning Analysis

For Bhili, Table 15 shows that, using XLS-R 300M, Hindi is the only donor that improves over the no-donor baseline in the reported condition. For Mundari, Hindi outperforms Santali in downstream ASR transfer, even though Santali is genealogically closer to Mundari. This supports the view that donor utility depends on acoustic representation compatibility, checkpoint quality, and fine-tuning stability, not family membership alone. Mundari donor candidates are Hindi, Santali, and Bengali.

Table 15: Donor-language initialization for Bhili using XLS-R 300M. *None* is the staged fine-tuning baseline at $\alpha=1.0$ (Table 13), and *Delta vs. Staged* gives each donor’s WER change relative to it. Lower is better; the best donor is shown in bold. Hindi is the only donor that improves on the no-donor baseline (−1.20 pp), whereas Marathi degrades it by 11.38 pp despite regional contact with Bhili-speaking areas.

Donor	WER (%)	Delta vs. Staged
None	46.53	—
Hindi	45.33	−1.20 pp
Marathi	57.91	+11.38 pp

H.6.1 Architectural Sensitivity to Transfer

For Whisper-Medium, transfer from Hindi hurts Bhili performance (42.03% vs. staged best 36.62%). CTC-based models such as wav2vec2 benefit from donor-language transfer, while seq2seq models such as Whisper already encode strong cross-lingual acoustic knowledge at scale. Additional transfer can therefore provide little

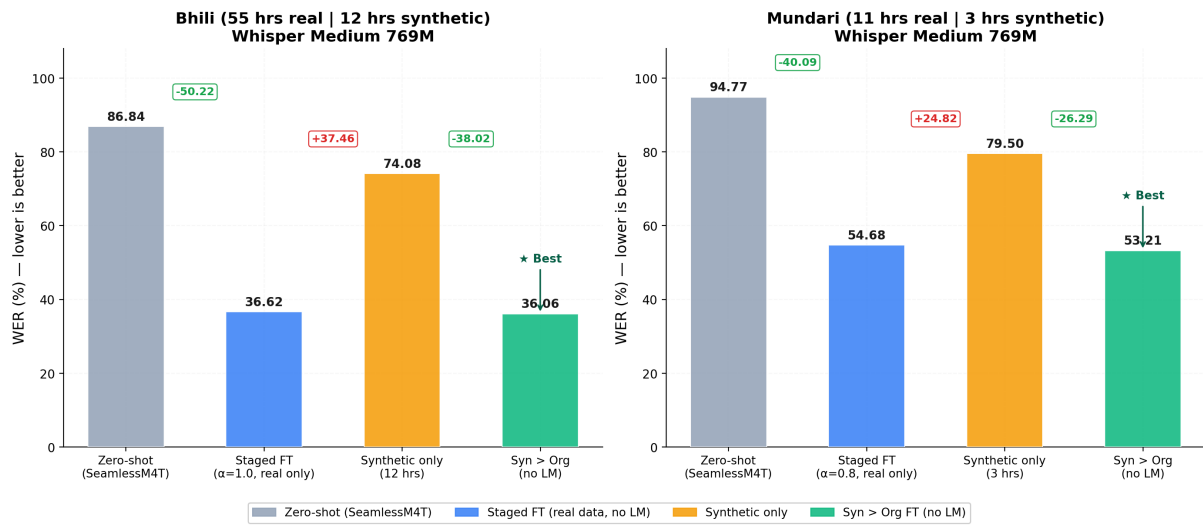


Figure 10: WER at each pipeline stage for Bhili (left) and Mundari (right) using Whisper-Medium; delta labels show per-transition WER change. Staged fine-tuning gives the largest single gain (-50.2 pp for Bhili; -40.1 pp for Mundari); synthetic pretraining helps Mundari and Bhili.

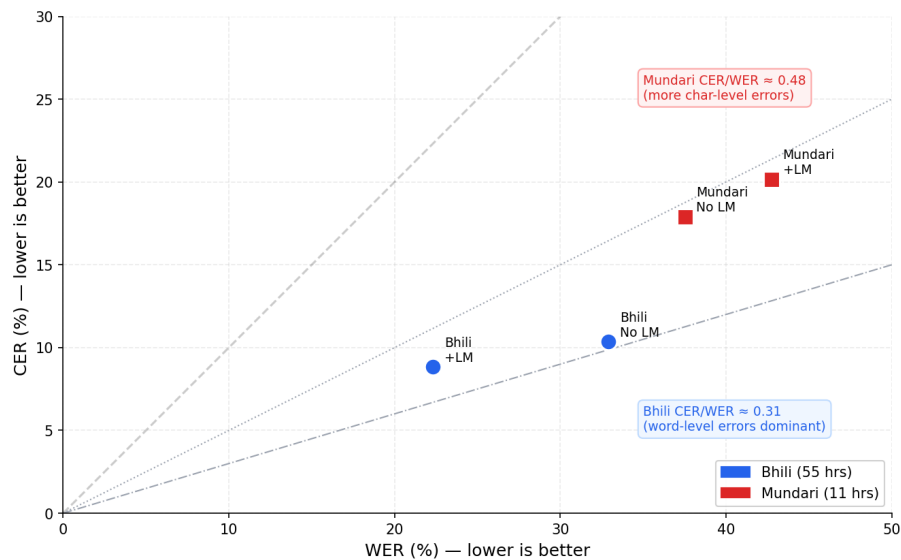


Figure 11: Post-LM WER versus CER for all configurations, with CER/WER ratio lines at 0.3 and 0.5 overlaid. Bhili systems cluster near 0.3 (word-boundary and lexical errors dominate, LM-correctable); Mundari systems cluster near 0.5 (persistent character-level acoustic confusion, requiring more speech data).

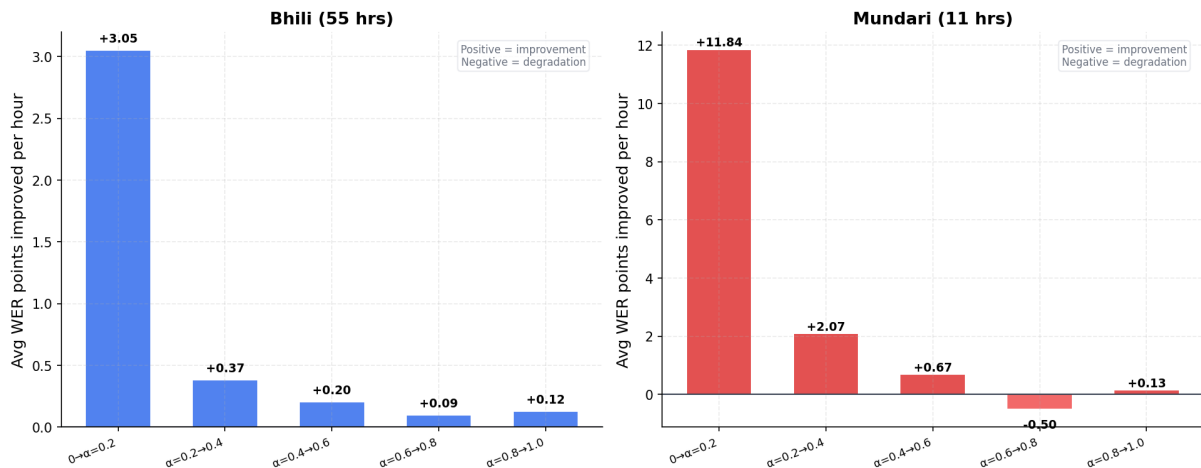


Figure 12: Average marginal WER improvement per effective training hour across α -transitions for Bhili (left) and Mundari (right); positive bars indicate gain, negative bars degradation. Bhili gains remain positive across 55h; Mundari plateaus after ~ 6 h and turns negative at $\alpha=1.0$.

benefit or even interfere. This architectural difference has practical implications for choosing a fine-tuning strategy.

H.7 Language Model Integration

We evaluate language-model integration for each language under two training paths, using Wav2vec2 xlsr-1B as the base acoustic model throughout. **Path A (Syn $\rightarrow$ Org+LM)** first trains on 12 h of synthetic speech for Bhili or 3h of synthetic speech for Mundari, then fine-tunes on the full real-data split and applies LM rescoring. **Path B (Original+LM)** uses only real-data fine-tuning before LM rescoring. Results are shown in Figures 13 and 3.

For Bhili, the **Original+LM** path is clearly strongest. It reaches 22.32% WER and outperforms Syn $\rightarrow$ Org+LM by 10.59 WER points, the largest gap between any two configurations in the experiment. The difference comes from how much each path benefits from LM rescoring: Original+LM improves by -21.74 pp ($44.06 \rightarrow 22.32$), a 49.3% relative reduction, whereas Syn $\rightarrow$ Org+LM improves by only -6.09 pp ($39.00 \rightarrow 32.91$), a 15.6% relative reduction.

This suggests that the acoustic model trained on 55 hours of real Bhili speech already provides a strong enough posterior for the LM to correct residual word-boundary and lexical-selection errors. By contrast, the Syn $\rightarrow$ Org model appears to retain acoustic artifacts from synthetic pre-training, leaving confusions that the LM cannot fully repair. The CER results support this interpretation: Original+LM achieves 8.83% CER, compared with

10.35% for Syn $\rightarrow$ Org+LM.

For Mundari, the pattern reverses. The **Syn $\rightarrow$ Org+LM** path reaches 37.52% WER and outperforms Original+LM (42.74%) by 5.22 WER points. Both paths benefit from LM integration, but the gain is larger for Syn $\rightarrow$ Org+LM: -27.66 pp ($65.18 \rightarrow 37.52$), a 42.4% relative reduction, compared with -19.27 pp ($62.01 \rightarrow 42.74$), a 31.1% relative reduction for Original+LM. Although Syn $\rightarrow$ Org starts from a worse pre-LM WER, it gives the better post-LM result.

The Mundari result suggests that synthetic pre-training changes the hypothesis space in a way that helps LM rescoring. TTS-generated transcripts are likely lexically cleaner and closer to the LM training domain than spontaneous Mundari speech, so the LM can rerank them more effectively. This is especially useful when the acoustic model is highly uncertain, as in the 11-hour Mundari setting, where the high pre-LM WER leaves more room for the LM prior to pull hypotheses toward valid word sequences.

Taken together, the two languages show different uses of LM integration. For Bhili, with 55 hours of real speech, synthetic augmentation is unnecessary: Original+LM avoids TTS artifacts, operates on a cleaner acoustic posterior, and yields the largest LM gain in the study (-21.74 pp). For Mundari, with only 11 hours of real speech, the 3-hour synthetic stage provides vocabulary bootstrapping and a cleaner hypothesis distribution for rescoring. The final 37.52% WER, compared with 42.74% for Original+LM, shows that in

Table 16: Effect of KenLM language model integration on Bhili and Mundari ASR performance. Table shows WER reduction (Δ WER) after LM rescoring for both training paths. For Bhili, the Original+LM path achieves 22.32% WER with substantial -21.74 pp improvement, outperforming Syn $\rightarrow$ Org+LM by 10.59 pp. Lower CER/WER ratios indicate word-level rather than character-level errors. All models are Wav2vec2 xlsr-1B, Ratio: CER/WER, B: Bhili, M: Mundari.

Lang	Path	WER(Pre)	WER(Post)	Δ WER	CER(Post)	Ratio
B	Syn $\rightarrow$ Org + LM	39.00	32.91	-6.09 pp	10.35	0.31
B	Original + LM	44.06	22.32	-21.74 pp	8.83	0.40
M	Syn $\rightarrow$ Org + LM	65.18	37.52	-27.66 pp	17.91	0.48
M	Original + LM	62.01	42.74	-19.27 pp	20.17	0.47

sub-15h conditions synthetic pre-training can help when followed by LM integration, even though Syn $\rightarrow$ Org alone (65.18%) is worse than staged FT alone (62.01%).

H.8 WER vs. CER Gap: Morphological Complexity

The CER/WER ratio helps separate word-level errors from deeper character- or phoneme-level confusion. A low ratio suggests that many word errors involve small character changes, such as boundary errors, lexical substitutions, or orthographic variants. A higher ratio indicates that word errors are more often accompanied by substantial character-level disagreement.

Results are visualized in Figure 11. For Bhili, the Syn $\rightarrow$ Org+LM path has a low CER/WER ratio (0.31), indicating that most word errors involve relatively small character-level changes. The acoustic model captures much of the phonetic content, while the remaining errors are concentrated around word boundaries, lexical choice, or orthographic variants. The Original+LM path has a slightly higher ratio (0.40), but it is also the best absolute system, reaching 22.32% WER and 8.83% CER. Its larger LM gain (-21.74 pp) suggests that the real-data-only model produces cleaner hypotheses that the language model can correct more effectively.

For Mundari, both paths have higher CER/WER ratios (0.47–0.48), meaning that nearly half of the word-level error is reflected at the character level. This points to a more acoustically difficult error profile. Mundari’s phonological and morphological structure, together with the limited 11-hour training set, leaves confusions that LM rescoring can reduce but not fully remove. Even so, the LM provides large absolute gains (-19 to -28 pp), because the high pre-LM WER leaves substantial room for language-model correction.

This decomposition helps identify whether future gains are more likely to come from better acoustic modeling or stronger language modeling. We treat **CER/WER** < 0.40 as an LM-friendly regime, where boundary and lexical errors dominate, and **CER/WER** > 0.45 as an acoustics-dominated regime, where more real speech, better donor selection, or phone-level transfer is likely to yield higher returns than LM tuning alone.

H.9 Data Efficiency: Marginal Return per Hour

Figure 12 compares the marginal WER reduction obtained from each additional block of effective target-language data. The two languages show different data-efficiency regimes. For Bhili, the largest gains occur early, especially between $\alpha = 0.2$ and $\alpha = 0.4$, where the model moves from 11 to 22 effective hours and average WER drops sharply. After about 33 effective hours, the gains become smaller but remain positive, indicating that the system has not saturated within the available 55 hours.

For Mundari, the first few hours matter most. The transition from zero-shot performance to 2.2 effective hours gives the largest marginal improvement, reducing WER from roughly 90% to the first fine-tuned operating point. After about 6 effective hours, the gains flatten, and some models begin to degrade. This pattern suggests that, in the 11-hour setting, additional target-language weight can increase overfitting unless it is balanced by donor-language mixing or other regularization.

These trends have direct implications for data collection. For a language with no transcribed speech, even 5–10 hours can produce a large reduction in WER. For Mundari, collecting another 20–30 hours would likely stabilize fine-tuning, reduce overfitting, and make the Syn $\rightarrow$ Org pipeline more reliable. For Bhili, the scaling curve has not

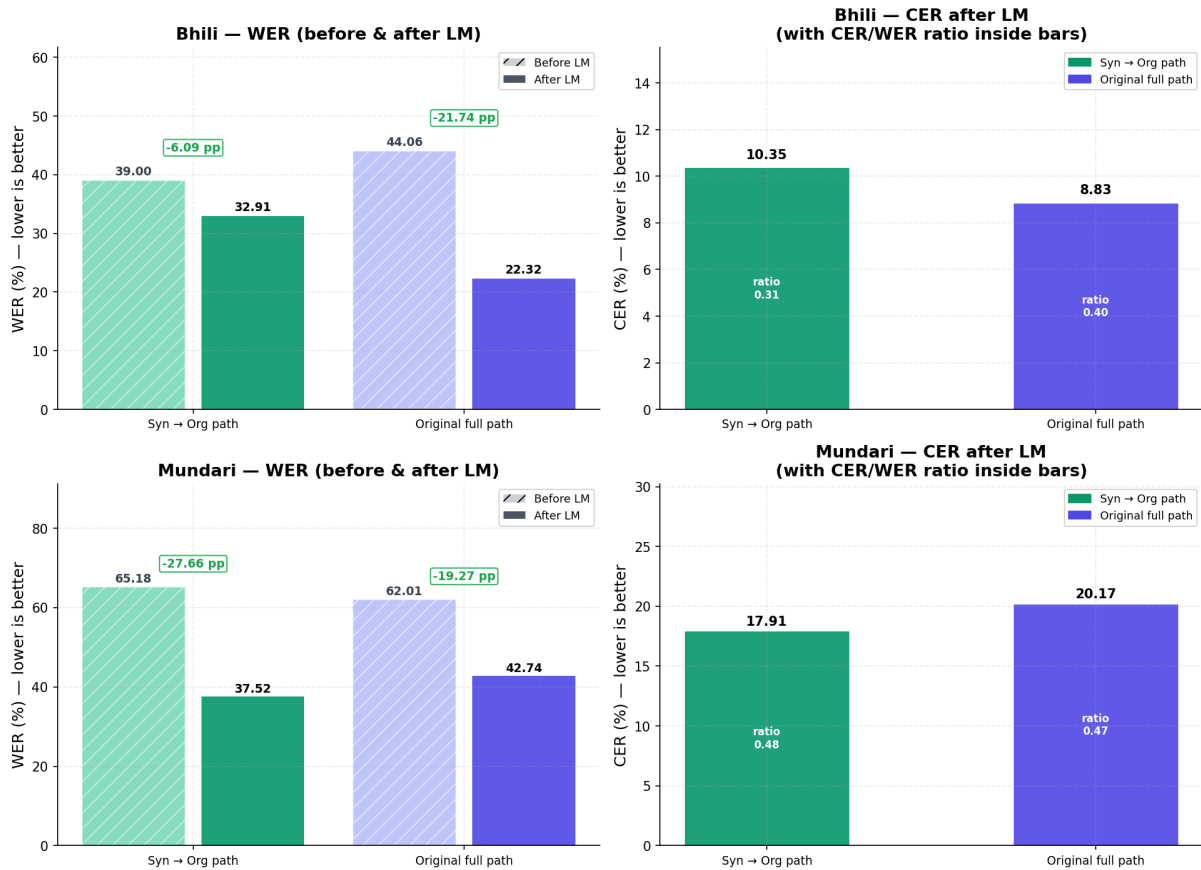


Figure 13: LM integration results for Bhili (top row) and Mundari (bottom row) across both training paths. Left panels show WER before and after LM for each path; right panels show post-LM CER with the CER/WER diagnostic ratio annotated inside each bar. Hatched bars indicate pre-LM values; solid bars indicate post-LM values.

yet flattened; extending the corpus toward 80–100 hours is likely to provide further gains and may push WER toward the 15–20% range under the current power-law extrapolation.

H.10 Best System per Configuration

The strongest configuration is language-dependent. For Bhili, **Original+LM** performs best, which suggests that synthetic pre-training is unnecessary once enough real speech is available. For Mundari, **Syn→Org+LM** gives the best result, showing that a small synthetic stage can help in the lower-resource setting when it is followed by real-data fine-tuning and LM rescoring. The overall comparison is shown in Figure 3.

The component-level gains also differ by language. For Bhili, staged fine-tuning is the largest single contributor (−37.86 pp), and LM rescoring adds a further −21.74 pp. For Mundari, staged fine-tuning again provides the largest initial gain (−34.94 pp), but LM integration on the Syn→Org path is especially important, adding −27.66 pp and exceeding the gain from the Original path (−19.27

pp).

At the best operating points, Bhili reaches 22.32% WER and Mundari reaches 37.52% WER, leaving a 15.2 pp gap. The gap is mainly explained by data quantity: Bhili has $5\times$ more real speech, which supports a stronger acoustic model and gives the LM a cleaner hypothesis space. The two languages also require different pipelines. Bhili performs best without synthetic data, whereas Mundari benefits from synthetic pre-training because only 11 hours of real speech are available. Finally, Mundari remains harder acoustically: its CER/WER ratio (0.47–0.48), compared with Bhili’s (0.31–0.40), indicates persistent character-level confusion that LM rescoring cannot fully remove.

H.10.1 Language Contrast: Bhili vs. Mundari

At their best configurations, Bhili reaches 22.32% WER and Mundari reaches 37.52% WER, leaving a 15.2 pp gap. This gap reflects three interacting factors:

Table 17: Post-LM WER, CER, and CER/WER ratio for Bhili and Mundari under both training paths. Bhili’s lower ratios (0.31–0.40) indicate LM-amenable lexical errors; Mundari’s higher ratios (0.47–0.48) indicate persistent acoustic confusion that additional language modelling alone cannot resolve.

Language	Path	WER	CER	CER/WER
Bhili	Syn→Org + LM	32.91	10.35	0.31
Bhili	Original + LM	22.32	8.83	0.40
Mundari	Syn→Org + LM	37.52	17.91	0.48
Mundari	Original + LM	42.74	20.17	0.47

Table 18: Incremental WER gains along the best Mundari pipeline. Staged fine-tuning contributes the largest single reduction (−34.94 pp); KenLM rescoring on the Syn→Org path adds a further −27.66 pp, totalling a 58.1% relative improvement from zero-shot to the best system.

Step	WER (%)	Gain
Zero-shot (Whisper-large-v3)	89.62	–
Staged FT ($\alpha = 0.8$, Whisper-Medium)	54.68	−34.94 pp
Syn→Org FT (Wav2vec2-1B)	65.18	(parallel path)
+ LM (Syn→Org+LM path)	37.52	−27.66 pp
Total from zero-shot	37.52	−52.10 pp (58.1% relative)

1. **Data quantity (primary cause):** Bhili has $5\times$ more real speech, which supports a stronger acoustic model, gives the LM a cleaner hypothesis space, and produces a −21.74 pp LM gain. Mundari obtains a larger absolute LM gain (−27.66 pp), but starts from a much higher pre-LM WER.
2. **Optimal pipeline divergence:** Bhili performs best without synthetic data (Original+LM), whereas Mundari benefits from synthetic pre-training (Syn→Org+LM). This reflects the 11h real-data bottleneck for Mundari.
3. **Phonological complexity:** Mundari’s CER/WER ratio (0.47–0.48), compared with Bhili’s (0.31–0.40), indicates persistent acoustic confusion that 11h of data and LM rescoring cannot fully resolve.

(1.61 pp) indicates that most of these errors are *aligned substitutions*: the Hindi cognate appears directly at the position of the Bhili reference token, rather than occurring incidentally elsewhere in the output. This points to a systematic dominant-language prior in the acoustic or language model, rather than random recognition noise.

For Mundari, HNB = 21.24% and HNBA = 17.41%, with a larger HNB–HNBA gap of 3.83 pp. This wider gap likely reflects the greater typological distance between Mundari and Hindi (Austroasiatic vs. Indo-Aryan): Hindi cognates in Mundari hypotheses are somewhat more likely to arise from incidental acoustic similarity than from direct aligned substitution. Even so, 17.41% of monitored tokens are replaced by Hindi forms at the aligned position, confirming a strong pretraining bias.

I Linguistic Metric Results and Analysis

I.1 Hindi-Normalisation Bias

Both best systems show substantial Hindi-normalisation bias despite competitive WER scores. For Bhili, HNB = 18.70% and HNBA = 17.09%, meaning that roughly one in five monitored-token positions is replaced by a Hindi cognate rather than a community-attested Bhili form. The small gap between HNB and HNBA

This pattern is not captured by WER or CER. The Bhili model reaches 22.32% WER but still Hindi-normalises nearly one in five monitored tokens. For language documentation, such substitutions are more harmful than many random errors: they replace Bhili material with fluent Hindi-like text, which can mislead downstream users and weaken corpus integrity.

Table 19: Linguistic evaluation metrics for the best model per language. HNB/HNBA: Hindi-Normalisation Bias (unaligned / alignment-constrained). MFR: Morphological Fragmentation Rate. SDR: Suffix Deletion Rate. All values are percentages.

Language	Model	HNB		Morphology		WER
		HNB	HNBA	MFR	SDR	
Bhili	xlsr-1B + KenLM (55h orig.)	18.70	17.09	0.62	3.40	22.32
Mundari	xlsr-1B + KenLM (3h syn.+11h orig.)	21.24	17.41	1.92	4.47	37.52

I.2 Morphological Integrity

Bhili (MFR = 0.62%, SDR = 3.40%). Morphological fragmentation is rare for Bhili, with 8 cases among 1,294 agglutinated tokens. This is consistent with Bhili’s relatively simpler suffixation system and with the 55 hours of training data, which provides greater exposure to inflected forms. Suffix deletion is more common (44 cases, 3.40%), especially for nasal and vowel-final suffixes (e.g., $-\dot{m}$, $-u\dot{m}$, $-ye$) that are acoustically subtle and phonologically conditioned. These suffixes mark grammatical number and tense in Bhili; omitting them can produce ungrammatical outputs even when WER remains acceptable.

Mundari (MFR = 1.92%, SDR = 4.47%). Mundari has higher error rates on both morphology metrics. MFR = 1.92% (21 cases) reflects the decoder’s tendency to split agglutinated Mundari tokens at morpheme boundaries. For example, *teko* (te+ko) is output as two tokens, *te* and *ko*. This follows from limited exposure: with only 11 hours of real Mundari data, the model has too little evidence that these morpheme sequences function as a single phonological word. SDR = 4.47% (49 cases) is dominated by loss of visarga suffixes (e.g., $-ah$, $-eh$, $-ih$), which are obligatory in Mundari nominal and verbal paradigms but have no acoustic parallel in the Hindi pretraining data. The model systematically removes these endings, producing bare stems that are either ungrammatical or belong to a different morphological category.

Interaction with WER. Table 20 shows that the proposed metrics are not redundant with WER. Bhili has much lower WER than Mundari (22.32% vs. 37.52%), yet its HNB is only 2.54 pp lower. Conversely, Mundari has higher WER and also worse MFR and SDR, reflecting a qualitatively different error profile driven by data scarcity and morphological complexity, not acoustic confusion alone.

The near-convergence of HNBA across the two

languages ($\Delta = 0.32$ pp), despite a 15.20 pp WER gap, is especially important. Both systems suppress the Hindi pretraining prior to roughly the same degree at aligned positions, regardless of overall recognition quality. This suggests that dominant-language bias is partly inherited from the pretraining regime and is not fully removed by fine-tuning on 11–55 hours of target-language speech, even when WER improves substantially.

I.3 Implications for Low-Resource ASR Development

The metric analysis suggests three practical recommendations:

1. **Collect more morphologically diverse training data.** SDR is the main morphological failure in both languages. Targeted collection of utterances containing inflected forms, especially acoustically subtle suffixes, is likely to reduce SDR more efficiently than increasing total audio hours uniformly.
2. **Address HNB at the language model level.** The small HNB–HNBA gap for Bhili suggests that Hindi cognates enter mainly through aligned acoustic confusion rather than LM hallucination. For Mundari, the larger gap suggests that the LM may sometimes promote Hindi cognates. Vocabulary-constrained decoding that penalises out-of-vocabulary Hindi tokens could reduce HNB without changing the acoustic model.
3. **Report HNB, MFR, and SDR alongside WER in future benchmarks.** WER alone is not sufficient for languages under dominant-language pressure. The near-identical HNBA scores despite a 15.20 pp WER gap show that WER-based model selection does not necessarily minimise documentation-harm errors. We recommend reporting HNB, MFR, and SDR as standard supplementary metrics for low-resource Indic ASR evaluation.

Table 20: WER and preservation metric scores for Bhili and Mundari, with per-metric deltas. Despite a 15.20 pp WER gap, alignment-constrained Hindi-normalisation bias differs by only 0.32 pp (HNBA: 17.09% vs. 17.41%), showing WER-based model selection does not minimise documentation-harm errors.

Language	WER	HNB	HNBA	MFR	SDR
Bhili	22.32	18.70	17.09	0.62	3.40
Mundari	37.52	21.24	17.41	1.92	4.47
Δ (Mundari – Bhili)	+15.20	+2.54	+0.32	+1.30	+1.07

J Linguistic Analysis of ASR Systems

We compare two low-resource ASR systems built on the XLS-R 1B architecture with a 5-gram KenLM language model: **Bhili**, fine-tuned on 55 hours of original speech, and **Mundari**, pre-trained on 3 hours of synthetic data, then fine-tuned on 11 hours of original speech. Both languages use the Devanagari script, enabling direct comparison of script-level error patterns.

J.1 Overall Performance

Bhili outperforms Mundari by 15.2 pp in WER and 9.08 pp in CER (Table 21, Figure 14), directly reflecting the $5\times$ difference in training data. Bhili’s utterance WER distribution is right-skewed with a mode near 0–5% and a small mean–median gap (3.9 pp), indicating a stable system. Mundari’s distribution is flatter and broader, with a larger mean–median gap (9.1 pp) caused by a heavy tail of severely erroneous utterances.

J.2 Error Type Distribution

Substitutions dominate in both models, but less so for Mundari (69.3%) than Bhili (79.2%). Mundari’s deletion rate (19.3%) is nearly double Bhili’s (10.1%), indicating the model frequently fails to hypothesize entire words, consistent with vocabulary gaps arising from the small training corpus.

J.3 Character-Level Error Analysis

Figure 15 shows the top-30 character substitution pairs and Figure 16 the full character confusion matrices.

Bhili The most frequent error is nukta loss: ज़→ज ($\times 29$) and ङ→ड ($\times 21$), where the diacritic distinguishing loanword phonemes is systematically dropped. Vowel-length matra confusions follow: ू→ु ($\times 23$) and ी→ि ($\times 20$). The sibilant merger श→स ($\times 14$) reflects a known Bhili dialectal feature.

Mundari Errors are more severe. The short/long vowel pair अ→आ ($\times 78$) is by far the most frequent single substitution — nearly three times the magnitude of Bhili’s top error. Nukta loss on ङ also occurs at high frequency ($\times 75$). The visarga ः is heavily confused with ा ($\times 18$) and आ ($\times 12$), reflecting a systematic failure to recognise the Mundari-specific final visarga morpheme, a symbol absent from Hindi and likely under-represented in XLS-R’s pre-training.

Figure 17 breaks substitutions by Devanagari character category. Consonant errors dominate in both models (Bhili: 517; Mundari: 849), followed by vowel matra errors (Bhili: 314; Mundari: 327). Mundari’s far higher consonant error count, despite fewer reference words, underlines its overall higher error density.

J.4 Word-Level Error Analysis

Bhili’s (Figure 18) top substitutions largely involve zero-width non-joiner (ZWNJ) omission, e.g. त्यो→त्यो ($\times 9$), an orthographic rather than phonetic error removable by Unicode normalisation. Mundari’s top pairs — ओड़ो:→ओड़ो: and रअ:→रअ: (both $\times 11$) — are similarly a normalisation artefact (ASCII colon vs. Unicode visarga in the reference). The genuinely phonetic pair तनअ:→तना ($\times 7$) illustrates simultaneous vowel and visarga deletion.

For insertions, Bhili’s model over-generates the Hindi postpositions से ($\times 8$) and के ($\times 7$), revealing language-model bias. Mundari’s model both deletes and inserts the postposition रे ($\times 13$ deleted; $\times 3$ inserted), indicating systematic confusion over agglutinative postposition placement.

J.5 Phonetic Proximity of Substitutions

Bhili’s substitutions are on average closer to the reference (normalised edit distance 0.49 vs. 0.58), see Figure 19, meaning errors differ in fewer characters relative to word length. Both models produce a bimodal normalised-distance distribution: a

Table 21: Global ASR performance for the best Bhili and Mundari systems. Bhili outperforms Mundari by 15.2 pp WER and 9.08 pp CER, directly reflecting the $5\times$ difference in training data; Mundari’s larger mean–median WER gap (9.1 pp vs. 3.9 pp) indicates a heavier tail of severely misrecognised utterances.

Model	WER	CER	Mean WER	Median WER	Total Errors
Bhili	22.32%	8.83%	20.0%	16.67%	1,513
Mundari	37.52%	17.91%	36.41%	27.27%	1,735

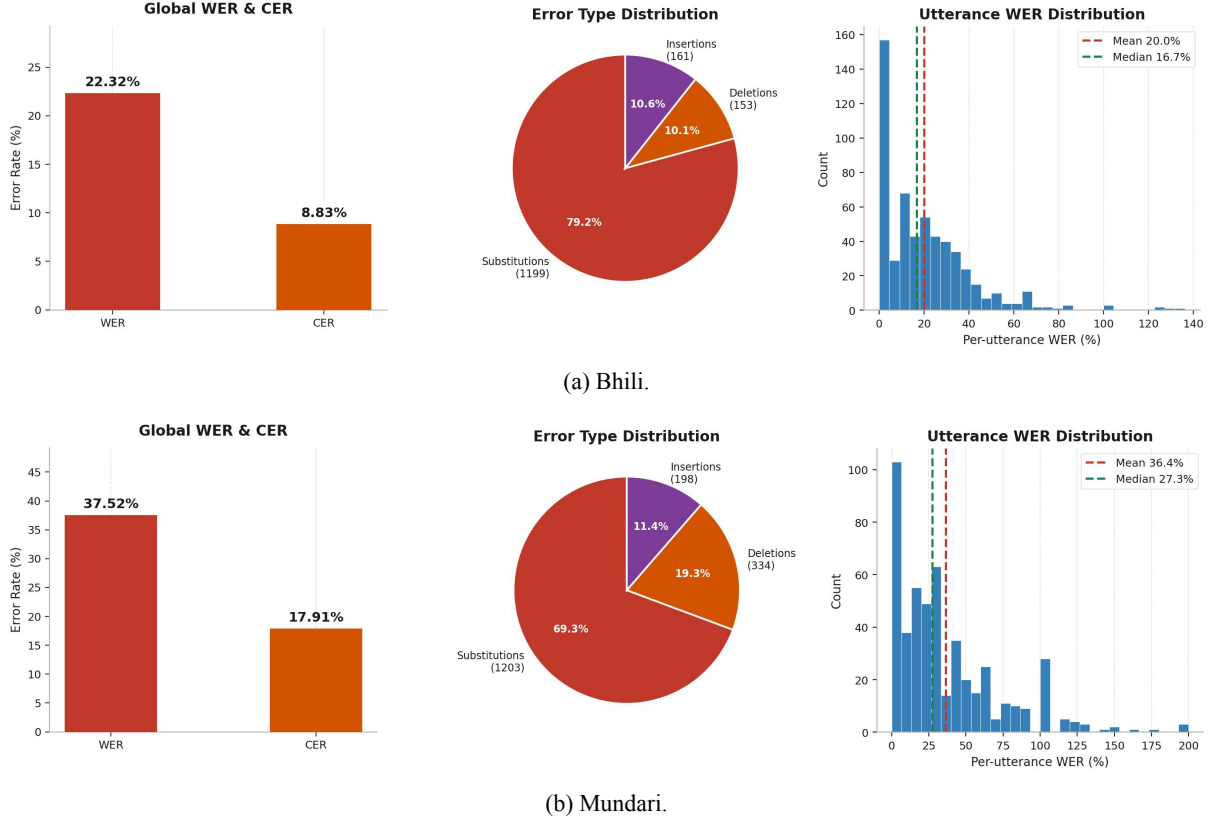


Figure 14: Global WER and CER, error-type breakdown, and per-utterance WER histogram for Bhili (top) and Mundari (bottom). Mundari’s deletion rate (19.3%) is nearly double Bhili’s (10.1%), and its broader WER histogram reflects systematic vocabulary gaps from the 11h training corpus.

near-zero peak (mostly ZWNJ/visarga normalisation errors) and a peak at 1.0 (completely wrong words), the latter more pronounced for Mundari.

J.6 Positional & Length Effects

Bhili’s error rate is essentially flat across the utterance (± 1.2 pp), indicating that the decoder maintains consistent performance. Mundari’s rate decreases from beginning to end. Models show flat and decreasing WER with utterance length (Figure 21) respectively, though Mundari’s improvement is more pronounced (from $\sim 40\%$ to $\sim 20\%$), suggesting that longer utterances contain more repetitive vocabulary the language model can exploit as shown in Figure 20.

J.7 Error Severity

Bhili produces a perfect transcript for 27.6% of test utterances (154 of 558) and falls into the Poor or Very Poor bands for only 7.9% (44 utterances), as shown in Figure 22 and Table 22. For Mundari, both ends of the distribution worsen: 19.2% of utterances (96 of 500) are perfect, while 28.6% (143 utterances) are Poor or Very Poor. More than one in four Mundari utterances is therefore severely misrecognised, and the Very Poor band alone accounts for 9.6% against 1.3% for Bhili. This asymmetry is consistent with the heavier tail reported in Section J.1, and indicates that Mundari output requires substantial post-editing before documentation use, even where its aggregate WER (37.52%)

Table 22: Percentage of utterances per WER severity band for Bhili and Mundari. Bhili achieves 27.6% perfect transcripts with 7.9% Poor or Very Poor; Mundari has only 19.2% perfect and 28.6% Poor or Very Poor, limiting downstream documentation usability without post-editing.

Severity band	Bhili	Mundari
Perfect (0%)	27.6%	19.2%
Excellent (<10%)	8.1%	3.2%
Good (10–30%)	39.6%	32.4%
Fair (30–50%)	16.8%	16.6%
Poor (50–100%)	6.6%	19.0%
Very Poor (>100%)	1.3%	9.6%

suggests a more uniform level of degradation.

K Training & Compute Infrastructure

K.1 Compute Infrastructure.

All experiments were conducted on NVIDIA A100 (40 GB) GPUs using the HuggingFace Transformers library. Mixed-precision training (FP16) was enabled throughout to reduce memory footprint. Training time ranged from approximately 4 to 18 GPU-hours per run depending on model size and dataset scale. All runs used the same hardware allocation to ensure fair comparison across configurations.

K.2 Hyperparameters

Table 23 summarises the training hyperparameters for each model family. All hyperparameters were fixed uniformly across languages and configurations to prevent any model from receiving a tuning advantage.

Wav2vec2 xlsr. The CTC-based xlsr-300M and xlsr-1B models were fine-tuned with the feature encoder frozen throughout training, following standard practice. A language-specific vocabulary was constructed from the training transcripts, retaining only Devanagari Unicode characters (U+0900–U+097F) plus a word-delimiter token. Text was NFC-normalised and all non-Devanagari characters were removed before vocabulary extraction. The CTC loss was computed with mean reduction, and decoding used greedy argmax over the output logits.

Whisper. The Whisper models were fine-tuned end-to-end without freezing any component. The language token was set to the closest supported

language (hindi) to initialise the decoder with a relevant multilingual prior; the task token was set to transcribe. Forced decoder ids were disabled during fine-tuning to allow the model to adapt its token distribution freely. Evaluation used `predict_with_generate` with beam search (beam width 5) and a maximum generation length of 448 tokens. Model cache was disabled during training for compatibility with gradient checkpointing.

Staged fine-tuning. For both model families, staged fine-tuning used a linear interpolation of target-language, controlled by the mixing ratio $\alpha \in \{0.2, 0.4, 0.6, 0.8, 1.0\}$, where $\alpha = 1.0$ corresponds to training on target-language data only. Donor-languages are used in full-split wherever needed. All other hyperparameters were held fixed across α values.

Language model integration. A 5-gram KenLM language model (Heafield, 2011) was trained on the target-language transcripts from the training split and applied at decoding time via shallow fusion. No separate hyperparameter sweep was performed for the LM weight; a fixed weight was used consistently across all configurations.

L Cultural Representation and Dataset Governance

AdiDhwani is designed around the principle that documentation-oriented ASR should preserve community-specific linguistic forms, not only produce fluent transcripts. The recordings include local lexical items, ecological and agricultural knowledge, kinship expressions, oral narratives, and institutional vocabulary. These forms are often precisely the material that makes a speech archive valuable for future language documentation.

The evaluation therefore includes both standard ASR metrics and preservation-oriented diagnostics. For Bhili, contact with Hindi, Gujarati, and Marathi makes dominant-language normalisation an important risk; for Mundari, agglutinative morphology makes suffix preservation central to grammatical fidelity. We therefore report Hindi-Normalization Bias (HNB), alignment-constrained HNB (HNBA), Morphological Fragmentation Rate (MFR), and Suffix Deletion Rate (SDR), so that culturally and linguistically meaningful changes remain visible during evaluation.

Table 23: Training hyperparameters for XLS-R and Whisper models; all runs use AdamW with FP16 mixed precision on A100 GPUs. Hyperparameters are fixed uniformly across all languages and configurations to ensure fair comparison; the only architectural difference is CTC greedy decoding for XLS-R versus beam-search generation for Whisper.

Hyperparameter	Wav2vec2 xlsr	Whisper
Batch size (per device)	8	8
Gradient accumulation steps	2	1
Effective batch size	16	8
Learning rate	1×10^{-4}	1×10^{-5}
Warmup steps	500	500
Max gradient norm (clipping)	1.0	—
Training epochs	20	20
Eval & save strategy	per epoch	per epoch
Best checkpoint criterion	min WER	min WER
Checkpoints retained	best only	3
Gradient checkpointing	yes	no
Feature encoder frozen	yes (first 6 layers)	n/a
Decoding	greedy (CTC argmax)	beam search (beam 5)
Max generation length	—	448 tokens

Dataset governance follows the same principle. Each recording contains a target-language utterance followed by a Hindi translation; only the target-language segment is used for ASR, while the Hindi tier supports annotation, checking, and auditability. Transcripts are produced through a maker-checker-superchecker workflow involving community speakers and expert linguistic review. Fixed speaker-disjoint splits, deterministic normalization, and preservation-oriented evaluation make the benchmark reproducible while keeping community-specific forms central to system diagnosis.

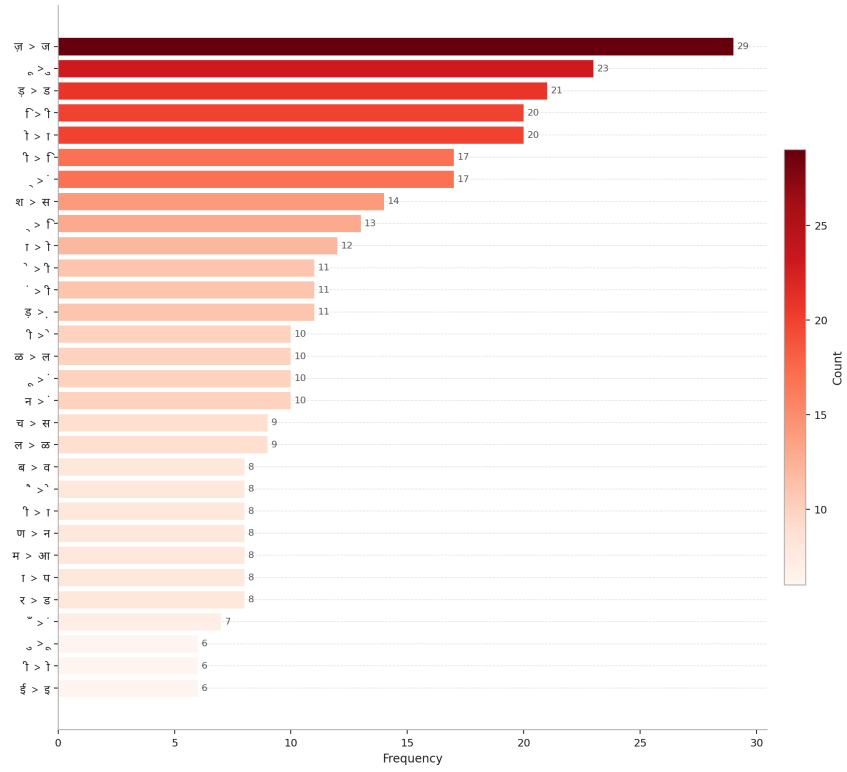
M Generalizing Beyond the Bhili-Mundari Benchmark

The transferable contribution of AdiDhwani is methodological. It provides a general evaluation recipe for documentation-oriented ASR: community-validated speech, fixed evaluation splits, reproducible adaptation baselines, and diagnostics that measure preservation-relevant errors alongside WER and CER.

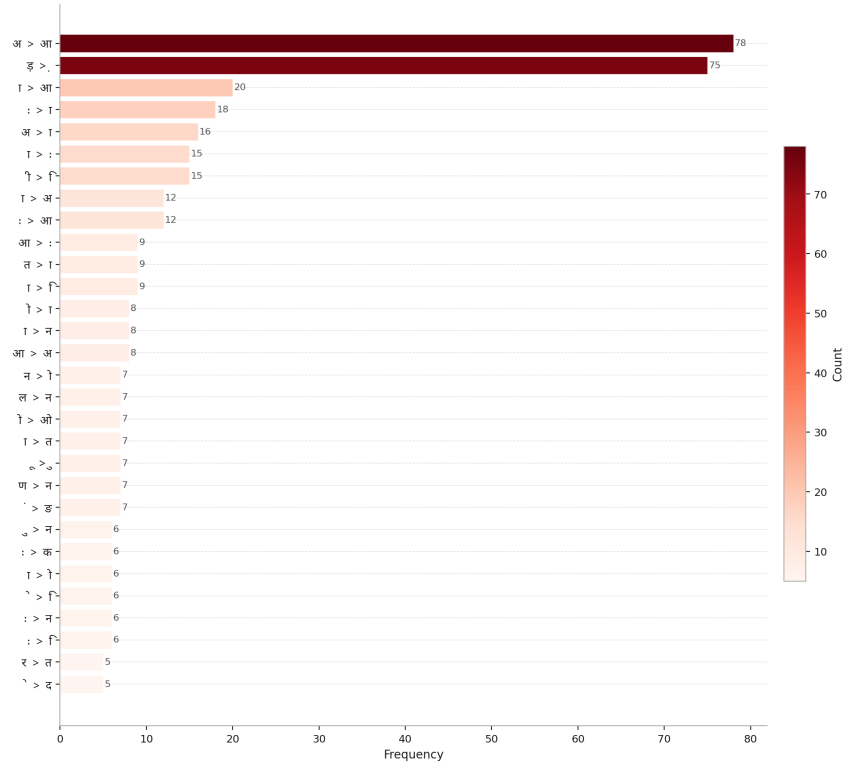
This framework can be applied to other under-documented languages by adapting the diagnostics to the community and language being studied. In some settings, the most important risks may be dominant-language normalisation; in others, they may involve morphology, tone, orthography, word segmentation, or named entities. The central principle is to evaluate not only whether a system recognises speech accurately, but also whether it preserves the linguistic forms that mat-

ter for documentation.

The broader lesson from AdiDhwani is therefore not a specific WER or CER value, but a model-development workflow: validate data with speakers, keep test sets fixed and speaker-disjoint, compare adaptation choices empirically, and report language-specific preservation diagnostics together with standard ASR metrics.



(a) Bhili.



(b) Mundari.

Figure 15: Top-30 character substitution pairs (reference → hypothesis) for Bhili (left) and Mundari (right). Bhili errors concentrate on nukta loss and vowel-length matrā confusion; Mundari shows severe short/long vowel confusion ($\times 78$) and systematic visarga deletion, a Mundari-specific morpheme absent from Hindi pretraining.

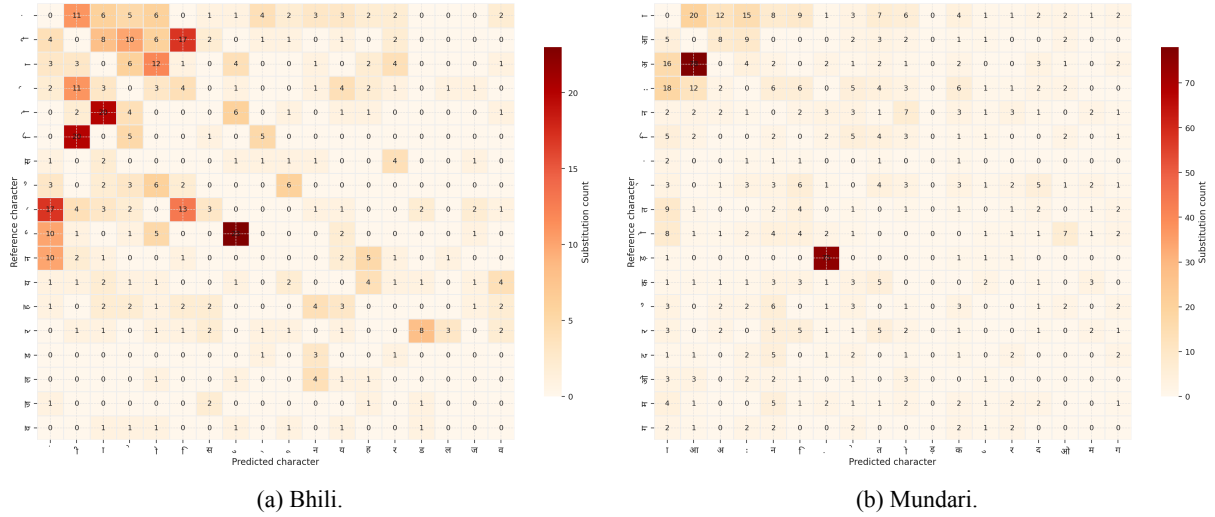


Figure 16: Character confusion matrices for Bhili (left) and Mundari (right); rows = reference, columns = hypothesis, cell intensity = substitution count. Bhili shows sparse near-diagonal errors; Mundari’s denser off-diagonal pattern reflects greater acoustic confusion from limited training data.

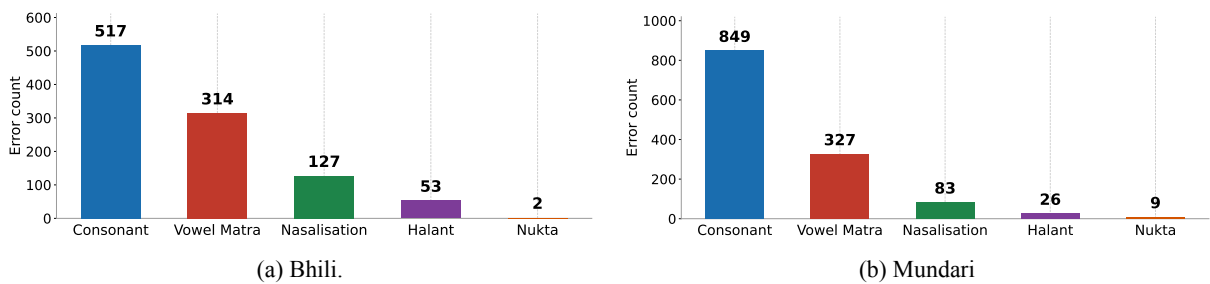
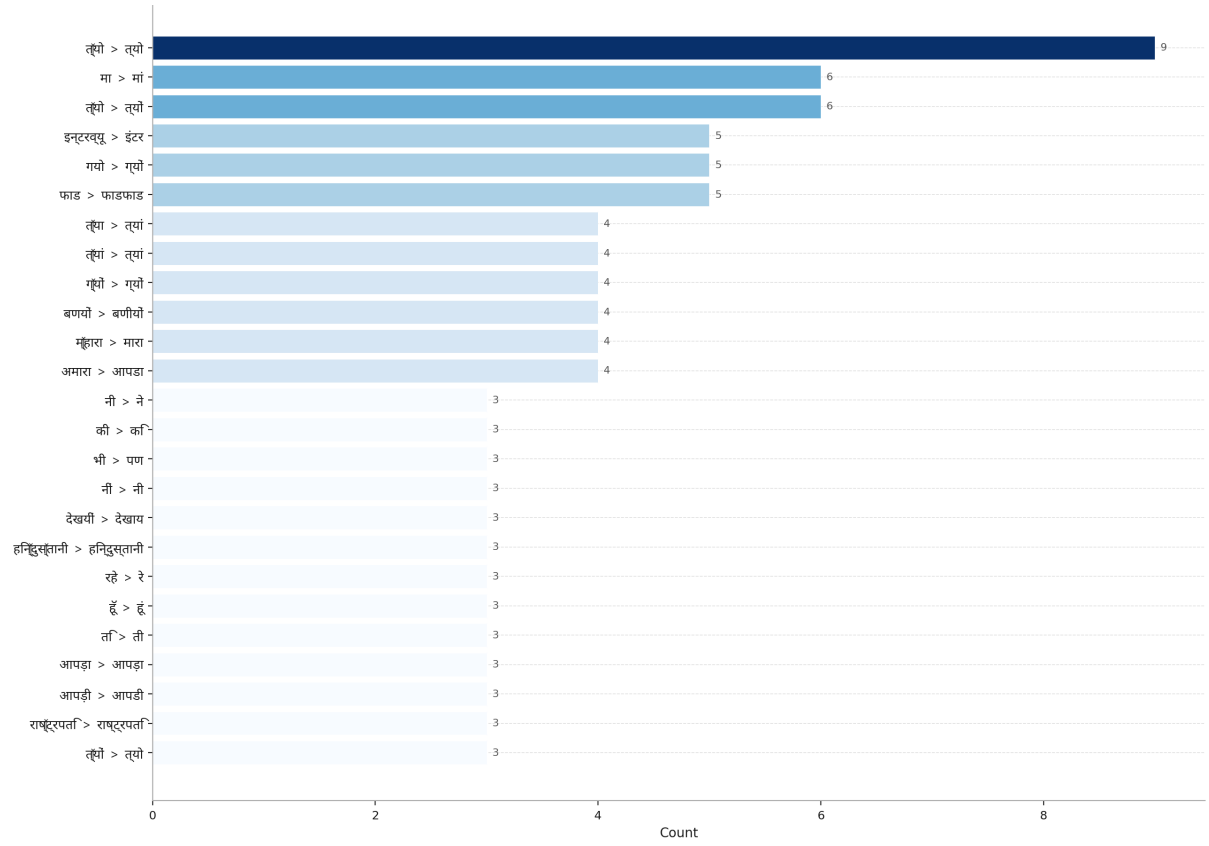
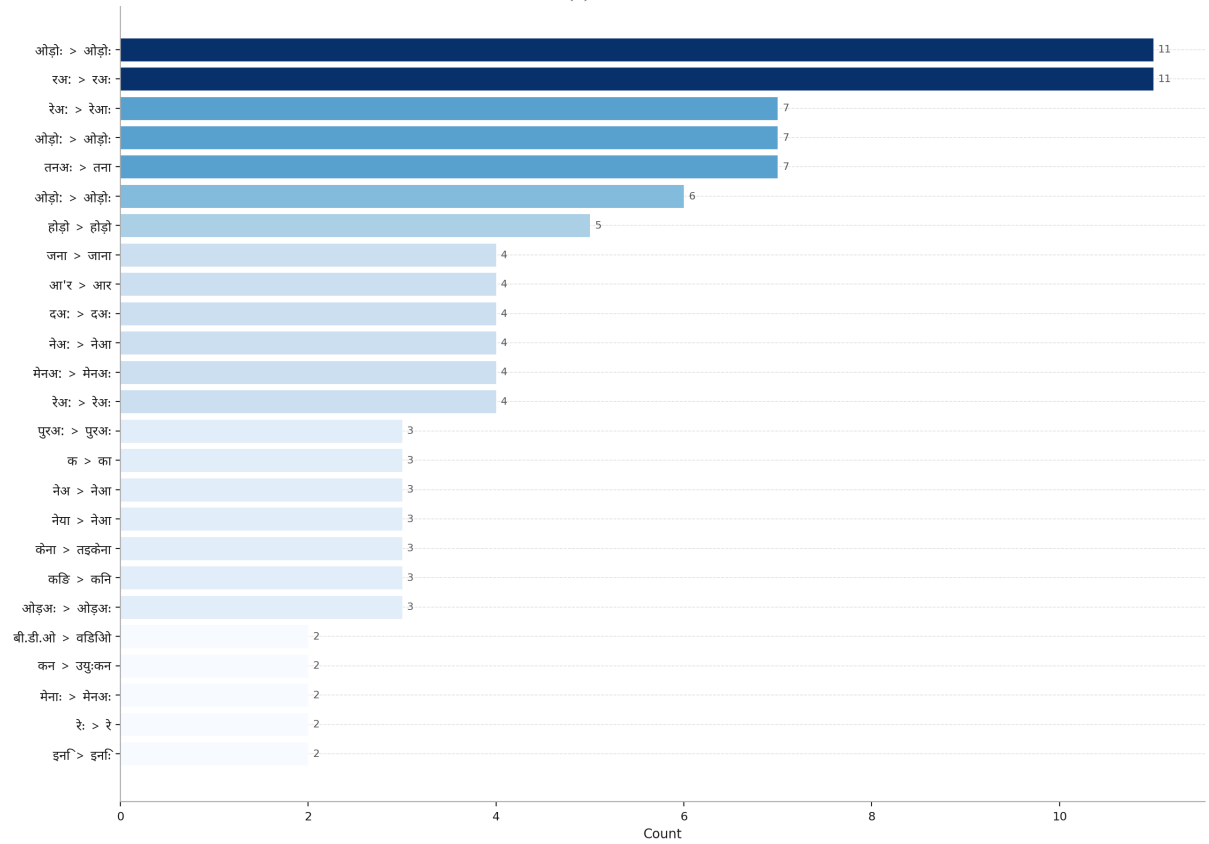


Figure 17: Substitution error counts by Devanagari character category for Bhili (left) and Mundari (right). Consonant errors dominate both languages; Mundari’s consonant error count (849) is $1.6\times$ Bhili’s (517) despite having fewer reference words, indicating higher per-word error density.

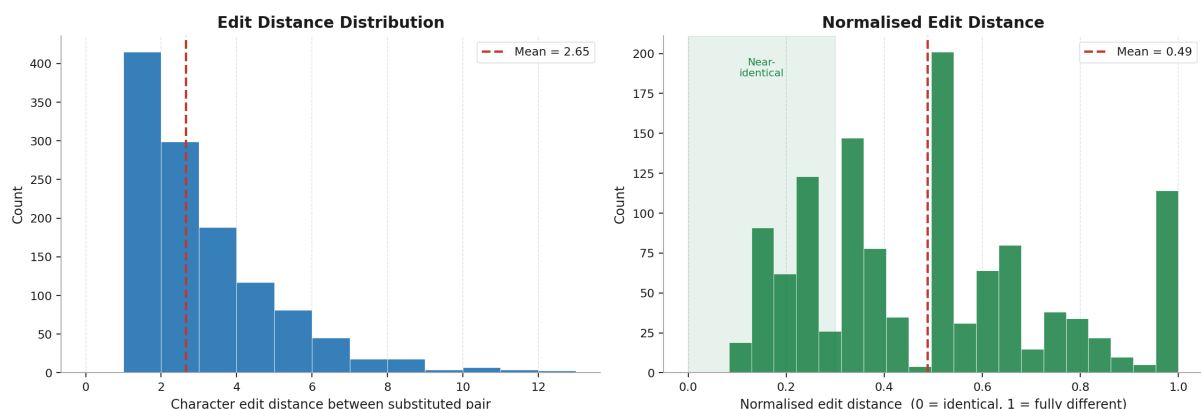


(a) Bhili.

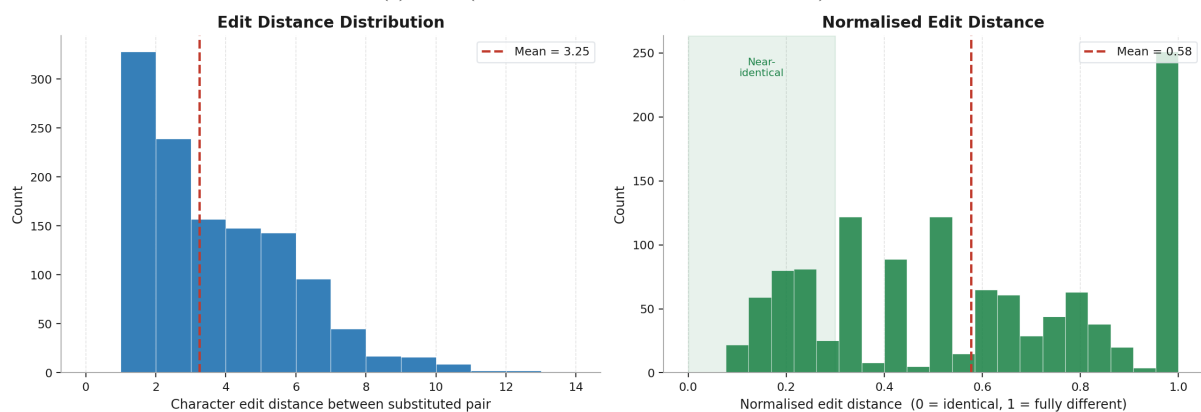


(b) Mundari.

Figure 18: Top-25 word substitution pairs for Bhili (left) and Mundari (right). Most Bhili errors are ZWNJ omission artefacts fixable by Unicode normalisation; Mundari shows genuine phonetic failures including visarga deletion. Bhili's model over-generates Hindi postpositions (से, के), directly evidencing Hindi-normalisation bias.

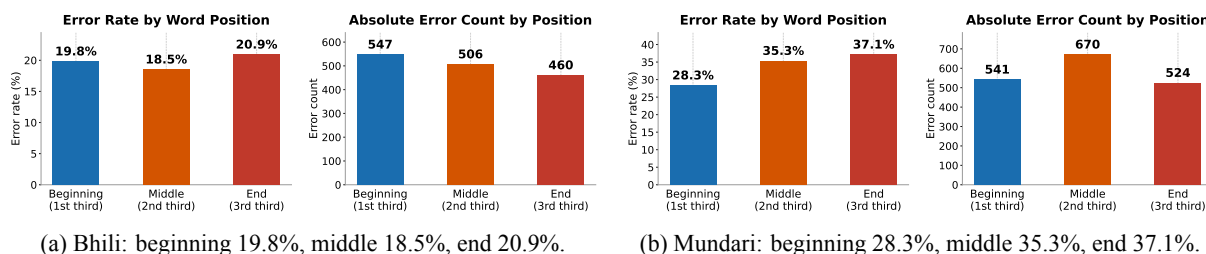


(a) Bhili (mean edit dist. = 2.65; norm. = 0.49).



(b) Mundari (mean edit dist. = 3.25; norm. = 0.58).

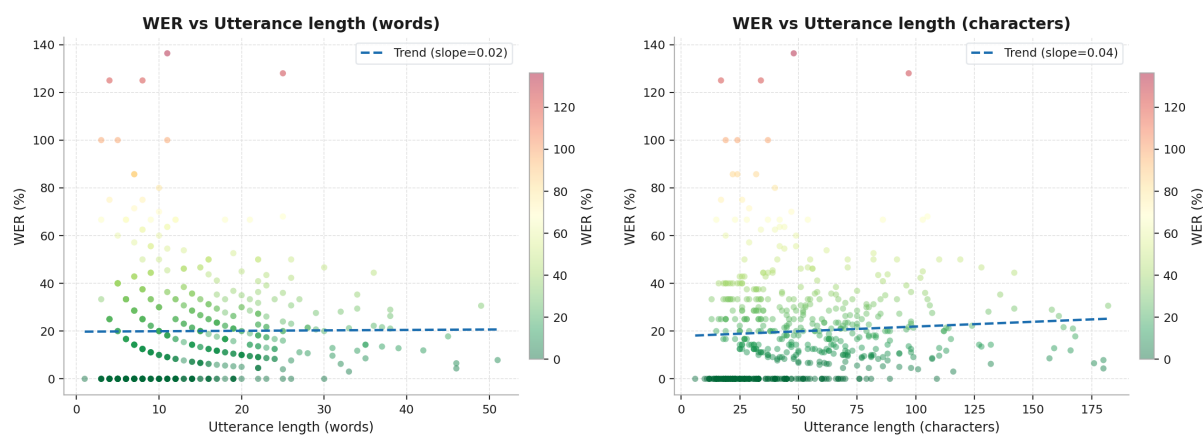
Figure 19: Character edit distance (left) and normalised edit distance (right) distributions for substituted word pairs in Bhili (norm. 0.49) and Mundari (norm. 0.58). Both show a bimodal pattern: near-zero errors are normalisation artefacts; the peak at 1.0 (complete misrecognition) is larger for Mundari.



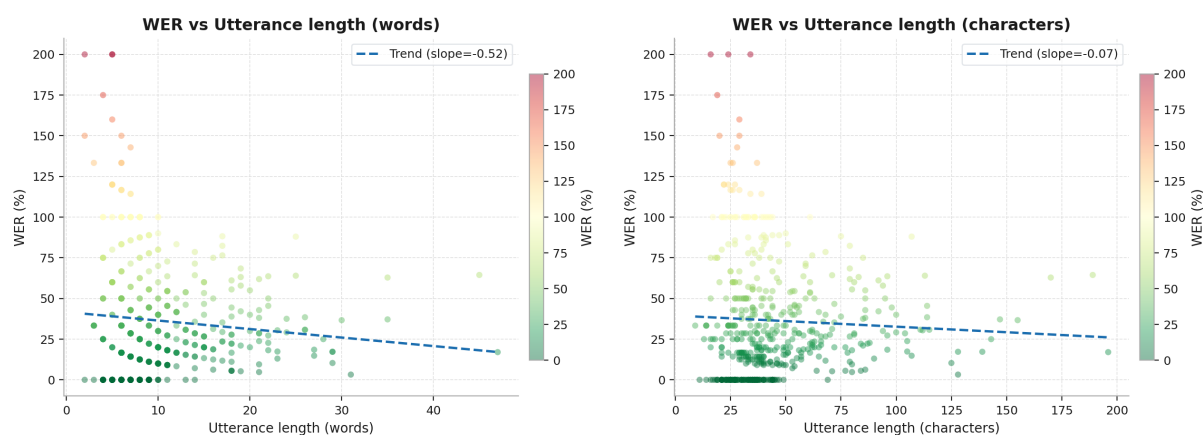
(a) Bhili: beginning 19.8%, middle 18.5%, end 20.9%.

(b) Mundari: beginning 28.3%, middle 35.3%, end 37.1%.

Figure 20: Error rate by utterance position (beginning, middle, end thirds) for Bhili (left; essentially flat at ± 1.2 pp) and Mundari (right; rising by 8.8 pp end-to-end). Mundari's increasing profile indicates the language model cannot suppress error accumulation as hypothesis length grows.

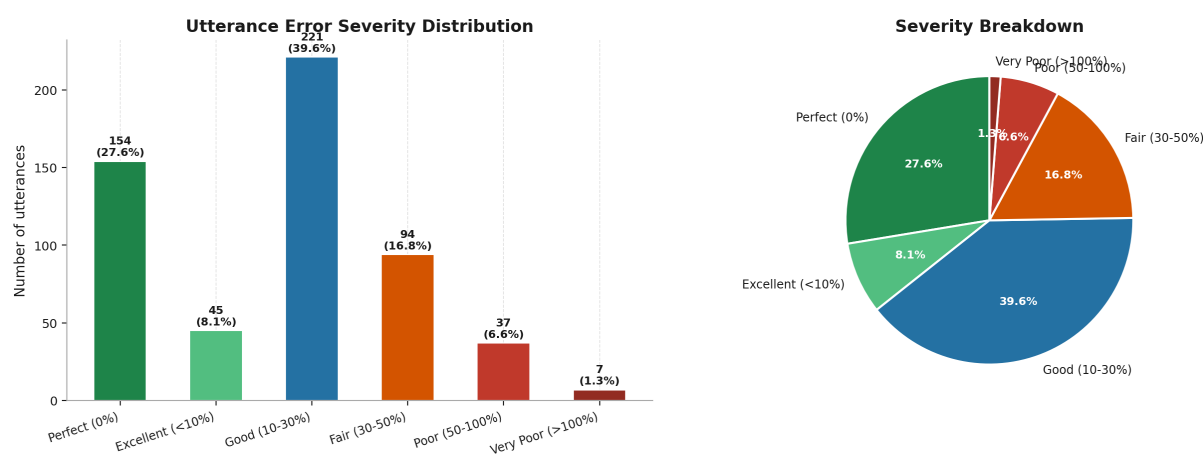


(a) Bhili.

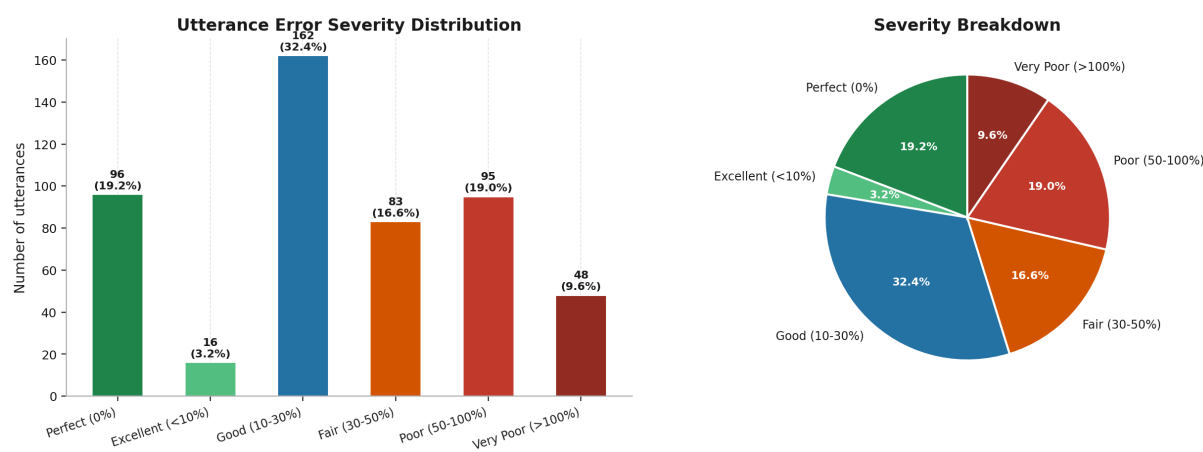


(b) Mundari.

Figure 21: WER versus utterance length in words (left) and characters (right) for Bhili (top) and Mundari (bottom). WER decreases with length for Mundari as the language model exploits repeated vocabulary; the effect is ($\sim 40\% \rightarrow 20\%$), reflecting greater LM reliance.



(a) Bhili



(b) Mundari

Figure 22: Per-utterance WER severity bands for Bhili (top) and Mundari (bottom); bar charts give utterance counts, pie charts the corresponding proportions. Bhili is perfect for 154 of 558 test utterances (27.6%) and Poor or Very Poor for only 44 (7.9%); Mundari is perfect for 96 of 500 (19.2%) and Poor or Very Poor for 143 (28.6%), so more than one in four Mundari utterances requires heavy post-editing for documentation use.