Dialect Matters: Cross-Lingual ASR Transfer for Low-Resource Indic Language Varieties

Akriti Dhasmana and Aarohi Srivastava and David Chiang
Computer Science and Engineering
University of Notre Dame
Notre Dame, IN, USA
{adhasman, asrivas2, dchiang} @nd.edu

Abstract

We conduct an empirical study of cross-lingual transfer using spontaneous, noisy, and code-mixed speech across a wide range of Indic dialects and language varieties. Our results indicate that although ASR performance is generally improved with reduced phylogenetic distance between languages, this factor alone does not fully explain performance in dialectal settings. Often, fine-tuning on smaller amounts of dialectal data yields performance comparable to fine-tuning on larger amounts of phylogenetically-related, high-resource standardized languages. We also present a case study on Garhwali, a low-resource Pahari language variety, and evaluate multiple contemporary ASR models. Finally, we analyze transcription errors to examine bias toward pre-training languages, providing additional insight into challenges faced by ASR systems on dialectal and non-standardized speech.

1 Introduction

Automatic speech recognition (ASR) systems for Indic languages have made significant progress through large-scale multilingual pre-training on high-resource, standardized languages. However, these advances have largely overlooked the linguistic reality of the region, where speech is characterized by extensive dialectal variation, spontaneous and noisy recording conditions, and frequent code-switching. As a result, state-of-the-art Indic ASR models often perform poorly when applied to low-resource dialects and language varieties, even those that are closely related to languages seen during pre-training.

Such efforts for Indic ASR typically rely on pre-training on a small set of mainstream languages such as Hindi, Marathi, and Bengali, under the assumption that phylogenetic similarity will facilitate transfer to related varieties. While this assumption provides a useful baseline in many cross-lingual settings, its sufficiency for dialectal speech remains underexplored. In practice, dialects and non-standard varieties often exhibit distinct phonological, lexical, and orthographic properties that are not well-captured by standard language data, raising questions about the effectiveness of relying solely on high-resource languages for transfer in dialectal ASR.

In this work, we examine cross-lingual transfer for Devanagari-script Indic varieties using spontaneous, noisy, and code-mixed speech. Rather than focusing solely on phylogenetic similarity, we investigate how different choices of fine-tuning data affect ASR performance on dialectal speech. Our results suggest that incorporating even limited amounts of dialectal speech during fine-tuning can be as effective as, or more effective than, relying on larger amounts of standardized language data, highlighting the role of dialectal variation in shaping transfer behavior.

To complement our cross-lingual analysis, we present a case study on Garhwali, a low-resource Pahari language variety that has been largely absent from prior ASR research. We evaluate several self-supervised speech models on Garhwali and perform a detailed error analysis to characterize the challenges posed by dialectal variation, code-mixing, and model bias toward pre-training languages. This case study provides concrete insights into the limitations of current Indic ASR models when applied to low-resource language varieties.

Contributions In this paper, we make three contributions:

1. We provide a comprehensive empirical analysis of cross-lingual transfer for low-resource Devanagari-script Indic dialects and show that fine-tuning on other dialects is often more effective than relying on closely-related high-resource standardized languages.

Proceedings of the Thirteenth Workshop on NLP for Similar Languages, Varieties and Dialects, pages 145–156
March 29, 2026. ©2026 Association for Computational Linguistics

2. We present the first detailed ASR study for Garhwali, including model evaluation and qualitative and quantitative error analysis.

3. We introduce a diagnostic approach for quantifying bias toward pre-training languages in dialectal ASR, enabling systematic analysis of model behavior on non-standardized and code-mixed speech.

2 Related Work

Recent work has highlighted systematic performance gaps between standard language varieties and regional or minority dialects across NLP and speech technologies. Kantharuban et al. (2023) provide a large-scale evaluation of state-of-the-art models for machine translation and automatic speech recognition across regional dialects of multiple high- and low-resource languages, showing that dialectal performance disparities are widespread and variably correlated with linguistic, social, and data-related factors. Complementary to this line of work, Blaschke et al. (2025) study standard-to-dialect transfer in spoken and written settings for German, demonstrating that speech-based models are more robust to dialectal variation than text-based or cascaded approaches, particularly when orthographic normalization is involved.

In the Indic context, prior work has investigated dialectal variation primarily through multilingual or language-specific ASR systems. For example, (Kumar et al., 2025) evaluate a multilingual dialect identification and ASR pipeline across 33 dialects of eight Indic languages using read speech, while other efforts have focused on building comprehensive ASR toolkits for standardized Indic languages (Singh Chadha et al., 2022). However, cross-dialectal and cross-lingual transfer for Indic ASRespecially in zero-shot settings and on spontaneous, non-standardized speechhas not yet been systematically examined. Our work addresses this gap by analyzing cross-lingual transfer behavior and dialectal bias in low-resource Indic language varieties.

The effects of orthographic irregularity and transcription variability on ASR performance have been examined in other language contexts, including cross-family settings (Taguchi and Chiang, 2024) and Swiss German dialectal ASR (Nigmatulina et al., 2020). We build on these findings to analyze how similar forms of orthographic variation impact ASR performance for Indic dialects.

Prior work on speech technology for Garhwali has primarily focused on language identification (Gusain et al., 2023) and the creation of domain-specific datasets, such as for agriculture-related applications (Riyal et al., 2016). To the best of our knowledge, no prior work has involved training or evaluating an ASR model for Garhwali.

The widespread presence of multiple language varieties in India, shaped by historical and sociolinguistic factors, has led to extensive code-mixing in everyday speech, posing additional challenges for ASR systems. While code-mixed speech has been studied in specific settings such as Hindi-Marathi ASR (Palivela et al., 2025), we provide a systematic way to quantify transcription errors arising from code-mixing in dialectal ASR.

3 Languages and Data

The Indic linguistic landscape is characterized by a high degree of diversity, encompassing hundreds of languages and dialects with varying levels of standardization and resource availability. Many widely spoken languages coexist with numerous regional dialects that differ substantially in phonology, morphology, and lexicon, despite phylogenetic relatedness (Massica, 1993). In everyday use, speakers frequently engage in code-mixing, particularly with English, and in spontaneous and acoustically noisy environments. These factors pose major challenges for ASR systems trained primarily on clean, standardized, and monolingual speech (Diwan et al., 2021). We present a phylogenetic tree of the languages included in our experiments in Figure 1.

Most existing ASR resources for Indic languages focus on a small subset of high-resource, standardized languages, leaving dialectal varieties underrepresented or entirely absent. To address this gap, we fine-tune and evaluate our models on the VAANI dataset (VAANI, 2025).[1] VAANI captures the rich linguistic diversity present across speakers of Indic languages, including different accents, grammatical variation, loan words, and code-switching patterns, all of which are vital to include in the context of Indic ASR.

VAANI consists of spontaneous speech collected by prompting participants to describe images in their local dialect. The dataset contains over 150,000 hours of speech, of which approx-

[1] https://huggingface.co/datasets/ARTPARK-IISc/VAANI

imately 10% is transcribed, and covers 156,534 speakers from 773 districts across India. As Indic languages are written in multiple scripts, we focus our analysis on language varieties written in the Devanagari script. Due to the broad coverage of dialects and language varieties, the amount of available data varies substantially across languages in the dataset; we account for this variability as much as possible in our experimental design.

Beyond conducting cross-lingual experiments across Indic language varieties, we also conduct a detailed analysis of ASR for one nonstandard language variety. We select *Garhwali* for this case study.[2] Garhwali is a language that belongs to the Pahari *subgroup* [3] spoken in the Himalayan region, particularly in the state of Uttarakhand (Gusain et al., 2023).

4 Experiments and Results

In our experiments, we aim to answer the following questions to analyze broader patterns in Indic ASR:

RQ1. How does a strong Indic ASR baseline perform across a diverse set of low-resource language varieties, particularly those not seen during model pre-training?

RQ2. To what extent does orthographic variability correlate with ASR performance in Indic language varieties?

RQ3. How does cross-lingual transfer for Indic ASR vary with phylogenetic distance, and does this relationship hold for dialectal speech?

In addition, we conduct a focused case study on Garhwali to characterize systematic error patterns in dialectal ASR. We analyze how systemic bias towards pre-training languages (Hindi) manifest in transcriptions generated by ASR models fine-tuned on a dialect (Garhwali).

Metrics We employ two standard metrics, word error rate (WER) and character error rate (CER), to evaluate the generated transcriptions. Both metrics are calculated based on Levenshtein distance, representing the minimum number of insertions, deletions, and substitutions required to align the hypothesis with the reference text. We report WER in our main results and include CER in the appendix.

4.1 RQ1: Baseline Assessment for Indic ASR

Setup We begin by assessing the performance of a state-of-the-art Indic ASR model, `IndicWav2Vec` (Kumar et al., 2022), on language varieties present in the VAANI dataset. `IndicWav2Vec` is based on the Wav2Vec 2.0 architecture (Baevski et al., 2020), and is pre-trained on 17,000 hours of unlabeled *clean* speech data from YouTube, as well as Newsonair data curated from radio channels, across 40 Indic languages. In our experiments, we employ `IndicWav2Vec-Hindi`[4] (`IndicWav2Vec` fine-tuned on Hindi), since pre-trained ASR models largely learn language-agnostic acoustic representations during pre-training and require language-specific fine-tuning to map these representations to text transcriptions.

Results We evaluate `IndicWav2Vec-Hindi` on test samples for all 30 Devanagari-script varieties in VAANI (up to 1 hour each) and report these results as part of a phylogenetic tree in Figure 1. We also divide the results into two tables depending on whether the language was used to pre-train `IndicWav2Vec` (Table 7) or it was not explicitly seen by the model (Table 8). These results show that, despite being fine-tuned on Hindi, `IndicWav2Vec-Hindi` achieves a best-case word error rate of 50.4% on VAANI Hindi test speech (one hour). Similarly high error rates are observed for several languages seen during pre-training, indicating that pre-training alone does not ensure robust performance on spontaneous and noisy speech, even for languages included in the pre-training corpus. We also observe substantial variation in performance across dialects and closely-related language varieties. These observations suggest that other factors, such as geographical proximity or phonological feature similarity, could have an impact on performance, motivating our subsequent analyses.

4.2 RQ2: Orthographic Consistency

Setup Orthographic consistency is an important factor in achieving high ASR performance, as inconsistent spellings introduce additional variabil-

[2] According to the Post-1971 Census of India, these languages were recognized as "mother tongues" and were designated as dialects of Hindi (Khubchandani, 1991); however, some linguists argue that Garhwali and other Pahari languages are distinct languages (Gusain et al., 2023).

[3] We refer to the lowest non-leaf nodes in a phylogenetic tree as *subgroups*.

[4] https://huggingface.co/ai4bharat/IndicWav2Vec-Hindindi

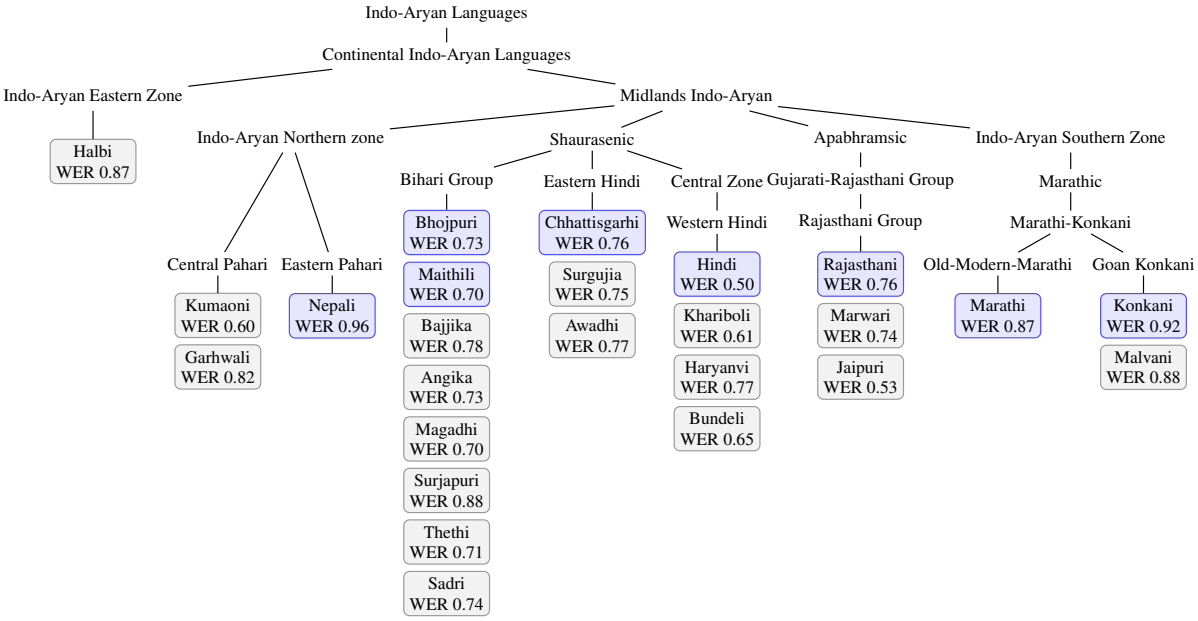

Figure 1: Subset of the Indo-European language family tree showing Devanagari-script Indic languages in the VAANI dataset, based on Glottolog (Hammarström et al., 2024). Languages are annotated with WER for `IndicWav2Vec-Hindi`. Blue highlights indicate languages used during pre-training.

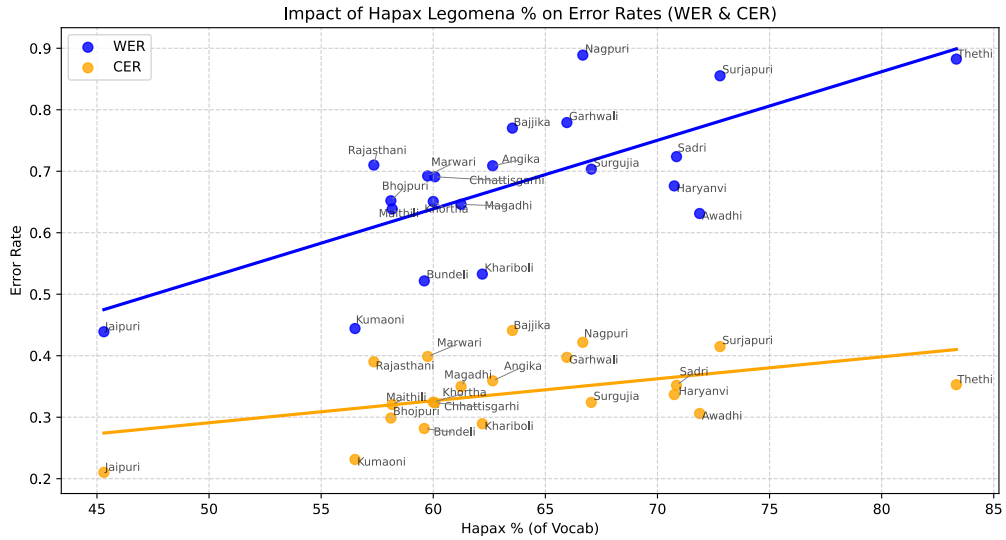

Figure 2: Hapax legomena (once-seen word types) % per language in the test split of VAANI plotted against WER (Pearson's $\rho = 0.705$, $p = 4 \times 10^{-4}$) and CER (*not significant*) using `IndicWav2Vec-Hindi`.

ity that can increase modeling and decoding errors (Nigmatulina et al., 2020). This challenge is particularly pronounced for Indic languages that encompass multiple dialects where orthographic conventions are not fixed. To quantify orthographic irregularity in the data, we measure the frequency of each word type in the transcripts and analyze the proportion of tokens that occur only once, along with the type-to-token ratio. We then examine how these measures correlate with ASR performance.

Results We observe a clear trend between orthographic variability, as measured by the proportion of *hapax legomena* (unique words) in the test set, and ASR performance. Figure 2 shows that languages with a higher percentage of unique words exhibit higher WERs when evaluated with `IndicWav2Vec-Hindi` (Pearson's $\rho = 0.705$, $p = 4 \times 10^{-4}$), indicating increased difficulty for languages with less consistent orthographic conventions (e.g., Thethi, Surjapuri). A similar but

weaker trend is observed for CER.

Additional evidence is provided by the type-token statistics reported in Table 9. Several languages (e.g., Gondi, Thethi) exhibit a high number of distinct word types relative to the total number of tokens, resulting in elevated type-to-token ratios. These measures reflect greater orthographic and lexical variability and are associated with higher ASR error rates. Taken together, these results suggest that orthographic inconsistency is an important contributing factor to reduced ASR performance in nonstandard Indic varieties.

4.3 RQ3: Cross-Lingual Transfer

We wanted to examine whether there are any performance gains from cross-dialectal fine-tuning in zero-shot settings. For each of these experiments, we fine-tune `w2v-bert-2.0`[5] (selected through preliminary comparison of models) on 1 to 7 hours of speech, depending on the availability. Each model was then evaluated on the test set of all Devanagari script languages available from VAANI. We compute correlations between the error rate and phylogenetic distance under two conditions: across all evaluation languages and restricted to nonstandard test varieties. We additionally measure the association between the error rate and whether the fine-tuning language represents dialectal or standard speech. We consider the following languages from the dataset to be dialects or nonstandard varieties: Angika, Awadhi, Bajjika, Bhili, Chhattisgarhi, Garhwali, Halbi, Haryanvi, Jaipuri, Khariboli, Khortha, Kumaoni, Magadhi, Malvani, Marwari, Nagpuri, Surjapuri, and Thethi.[6]

Results Figure 3 summarizes cross-lingual transfer performance for `w2vBERT` models fine-tuned on 1 to 7 hours of data per language and evaluated in a zero-shot setting across a wide range of Indic language varieties, providing a holistic view of how ASR performance varies as a function of both the fine-tuning and the evaluation languages.

Across all evaluation languages, we observe a clear association between phylogenetic distance and ASR performance: larger distances between the fine-tuning and evaluation languages generally correspond to higher word error rates (Spearman's $\rho = 0.333$, $p = 1.11 \times 10^{-7}$). This result aligns with prior findings in cross-lingual ASR and confirms that phylogenetic relatedness provides a useful baseline for transfer across Indic languages. However, this aggregate trend does not fully characterize model behavior when evaluation is restricted to dialectal speech.

4.3.1 Dialectal Transfer Effects

When focusing specifically on dialects and nonstandard test varieties, the heatmap reveals additional structure that is not explained by phylogenetic distance alone. Although the association between phylogenetic distance and word error rate remains statistically significant in this setting (Spearman's $\rho = 0.274$, $p = 3.604 \times 10^{-4}$), several consistent deviations from this trend are observed. In particular, models fine-tuned on nonstandard varieties frequently outperform models fine-tuned on phylogenetically closer but standard, higher-resource languages.

One representative example from Figure 3 is the strong transfer from Marwari (a dialect of Rajasthani) to Kumaoni (a Pahari language variety), even though the two belong to distinct phylogenetic subgroups and are geographical distant. Another example is the consistently competitive performance of the models fine-tuned on Marwari and Magadhi across multiple evaluation language varieties, often outperforming models fine-tuned on higher-resource standardized languages such as Hindi, Marathi, and Rajasthani. Notably, this trend holds despite the fact that the Magadhi model is trained on less data (5 vs. 7 hours). These cases illustrate that while fine-tuning on a closely related high-resource standardized language may be the most natural strategy, it is not necessarily the most effective choice for transfer to dialects and related language varieties.

To further examine this pattern, we analyze whether the dialectal status of the fine-tuning language itself is associated with performance on unseen dialects. We find a statistically significant trend indicating that fine-tuning on dialectal speech is associated with lower WER on dialectal evaluation sets (point-biserial correlation $r_{pb} = -0.196$, $p = 5.79 \times 10^{-3}$). This trend holds even when the available dialectal training data is smaller than that of the corresponding standardized languages, suggesting that dialectal fine-tuning captures information that is not adequately represented by mainstream language data alone.

[5] https://huggingface.co/facebook/w2v-bert-2.0

[6] None of these languages belong to the Eighth Schedule to the Indian Constitution listing the officially recognized languages (https://en.wikipedia.org/wiki/Eighth_Schedule_to_the_Constitution_of_India).

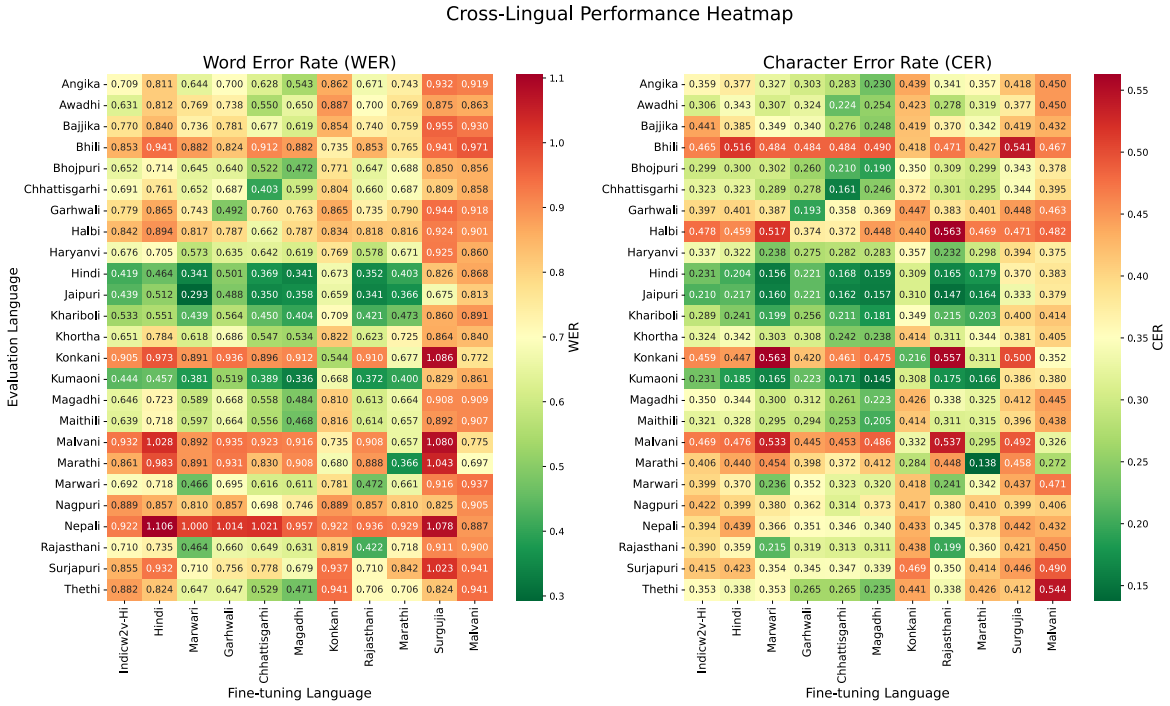

Figure 3: Cross-Lingual Performance of w2vBERT models fine-tuned on 1 to 7 hours of data per language vs. off-the-shelf IndicWav2Vec-Hindi.

Fine-Tuning Language	Hours	Jaipuri WER	Jaipuri CER
Rajasthani + Marwari	16.8	0.309	**0.151**
Marwari (*dialect only*)	7.2	**0.301**	0.162

Table 1: Effect of fine-tuning language choice on Jaipuri ASR (Rajasthani subgroup). Mainstream language is in bold.

Fine-Tuning Language	Hours	Thethi WER	Thethi CER
Angika, Bajjika, **Bhojpuri**, Khortha, Magadhi, **Maithili**, Sadri	50.6	0.412	**0.132**
Angika, Bajjika, Khortha, Magadhi, Sadri (*dialects only*)	15.6	**0.353**	0.147

Table 2: Effect of mainstream language inclusion on Thethi ASR (Bihari subgroup). Mainstream language is in bold.

Fine-Tuning Language	Hours	Haryvani WER	Haryvani CER
Hindi + Bundeli + Khariboli	101.7	0.930	1.452
Bundeli + Khariboli (*dialects only*)	1.7	**0.513**	**0.267**

Table 3: Impact of fine-tuning (FT) languages on Haryanvi ASR (Western Hindi subgroup). Mainstream language is in bold.

Controlled In-Group Transfer Experiments To isolate the effect of dialectal vs. standard fine-tuning data from broader cross-lingual trends, we conduct controlled transfer experiments within individual phylogenetic subgroups. These experiments compare models fine-tuned on dialectal varieties against models fine-tuned on higher-resource standardized languages within the same subgroup. Across the Rajasthani, Western Hindi, and Bihari subgroups, we consistently observe that fine-tuning on small amounts of dialectal speech yields performance comparable to or better than fine-tuning on substantially larger amounts of standardized language data. For example, within the Rajasthani subgroup (Table 1), a model fine-tuned solely on a dialectal variety (Marwari) achieves comparable performance on an unseen dialect (Jaipuri) to a model trained on a combination of dialectal and macro-language data. Similar trends are observed in the Western Hindi and Bihari subgroups (Tables 3 and 2), where excluding higher-resource mainstream languages does not degrade, and in some cases improves, performance on dialectal evaluation sets.

Taken together, the heatmap analysis and controlled in-group experiments demonstrate that while phylogenetic similarity provides a useful starting point for cross-lingual transfer, effective

Model	Metric Optimized	WER	CER
XLS-R	WER	0.650	0.270
wav2vec2-BERT	CER	**0.493**	**0.193**
Whisper-small	CER	0.629	0.650
HuBERT	CER	0.515	0.199

Table 4: Comparison of speech models fine-tuned on Garhwali. Training details are in Table 10.

Model	# Non-Hi	Correct	To Hi	To Wrong
wav2vec2-BERT	1873	34.8%	23.3%	38.8%
Indicw2v2	2109	4.2%	37.3%	42.9%

Table 5: Comparison of w2vBERT model fine-tuned on Garhwali against IndicWav2Vec-Hindi on Garhwali non-Hindi word handling. Key: Hi = Hindi.

ASR for dialectal speech depends critically on the inclusion of dialectal training data itself.

5 Garhwali ASR Case Study

5.1 Baseline Selection

Setup In order to determine the ideal model architecture for Garhwali, we compared several widely-used ASR models:

- Wav2Vec2 (Baevski et al., 2020)
- HuBERT (Hsu et al., 2021)
- XLS-R (Babu et al., 2022)
- Whisper (Radford et al., 2023)
- w2vBERT (Chung et al., 2021)

We fine-tune each of these models on the Garhwali subset of VAANI, using an 87%–6%–7% train–test–validation ratio.

We fine-tuned and evaluated several self-supervised speech models on the Garhwali subset of the VAANI dataset. Table 4 reports performance across these models.

Results Among the evaluated configurations, the w2vBERT-based model achieves the lowest error rates and is therefore selected for subsequent analysis. We compare this model against IndicWav2Vec-Hindi (used in our cross-lingual experiments). Our fine-tuned model performs better than the IndicWav2Vec-Hindi model; however, the resulting error rates remain high, indicating that Garhwali ASR remains challenging even with dialect-specific fine-tuning. We therefore focus our error analysis on this best-performing configuration to better understand the remaining sources of error.

5.2 Error Analysis

We conduct an extensive error analysis of the best fine-tuned model on Garhwali to identify the following:

Inconsistent English Transliteration English text is marked in VAANI's annotations; we assess ASR output specifically on these segments. A prominent source of error arises from inconsistent transliteration of English words into the Devanagari script. Due to frequent code-switching in spontaneous Indic speech, English words commonly appear in the ground-truth transcripts. However, the absence of standardized conventions for English-to-Devanagari transliteration, combined with imperfect phoneme-to-grapheme mappings, results in substantial variation in the labels. For example, the training data contains multiple spellings for the English word *photo*, such as फोटु and फाटो, reflecting pronunciation differences.

This variability introduces ambiguity during training and evaluation, complicating the model's ability to learn consistent lexical representations. Whether standardizing transliterated forms would improve dialectal ASR performance or instead remove linguistically meaningful variation remains an open question.

Bias Towards Hindi from Pre-training We identify non-Hindi words in the ground-truth transcripts using Hindi-HunSpell,[7] a Hindi spell-checker. These includes Garhwali words, as well as transliterated English words; we separate the English words (see below). We then assess which non-Hindi words are preserved in the generated transcription vs. which are transcribed to a Hindi word. We analyze commonly omitted, added, and substituted characters in the generated transcriptions, qualitatively identifying systematic errors and highlighting areas for future improvement. We also observe systematic bias towards Hindi arising from model pre-training. Many errors involve Garhwali words or transliterated English terms being incorrectly normalized to valid Hindi words.

To quantify this effect, we identify non-Hindi words in the ground-truth transcripts, including Garhwali-specific vocabulary and transliterated English terms, using a Hindi spell checker. We then track three outcomes: non-Hindi words cor-

[7] https://github.com/Shreeshrii/hindi-hunspell

Actual		Predicted		#	Actual		Predicted		#
Deva	IPA	Deva	IPA		Deva	IPA	Deva	IPA	
␣		ϵ		438	ϵ		ं	/n/	68
ϵ		␣		289	ु	/ʊ/	ू	/uː/	66
ा	/aː/	ϵ		175	ी	/iː/	ि	/ɪ/	55
्		ϵ		170	े	/eː/	ϵ		51
ϵ		्		163	ू	/ʊ/	ϵ		51
ं	/n/	ϵ		145	ि	/ɪ/	ϵ		48
ϵ		ा	/aː/	126	ह	/ɦ/	ϵ		46
ू	/uː/	ु	/ʊ/	94	ϵ		े	/eː/	45
ि	/ɪ/	ी	/iː/	90	ै	/ɛː/	े	/eː/	44
ϵ		य	/j/	81	ϵ		क	/k/	44
ी	/iː/	ϵ		76	ब	/b/	भ	/bʱ/	42
य	/j/	ϵ		76	र	/r/	ϵ		42

Table 6: Most frequent transcription errors for Garhwali ASR. Key: ␣ = space; ϵ = empty string; Deva = Devanagari script.

rectly preserved (`Correct`), non-Hindi words converted into valid Hindi words (`No-Hi→Hi`), and non-Hindi words converted into incorrect forms (`No-Hi→Wrong`). Note that the total number of non-Hindi words in the ground truth transcripts differs slightly between the two models due to preprocessing and tokenization; however, the observed trends are robust to these discrepancies.

As shown in Table 5, the `w2vBERT` model fine-tuned on Garhwali converts approximately 22.9% of non-Hindi terms into Hindi words, while preserving roughly one-third of such terms. In contrast, `IndicWav2Vec-Hindi` preserves fewer than 5% of non-Hindi words, converting or distorting the majority. These results indicate that ASR pre-training on mainstream Indic languages can substantially hamper a model's ability to retain dialect-specific information.

Character-Level Error Patterns We further analyze character-level errors produced by `w2vBERT` fine-tuned on Garhwali. Table 6 summarizes the most frequent substitutions, insertions, and deletions, which can be grouped into three broad categories: (1) word-boundary errors (e.g., space versus deletion), (2) vowel and consonant length confusions (e.g., ु vs. ू), and (3) aspiration-related errors (e.g., ब vs. भ).

Word-boundary errors are the most frequent, with 438 instances of omitted spaces and 289 instances of spurious space insertions, contributing to the observed WER despite relatively lower CER. The next most common errors involve vowel length and halant usage. In Garhwali, consonant-final words and consonant clipping, often marked using the halant ् , are common, whereas Hindi typically enforces an inherent vowel at word endings. Models pre-trained on Hindi orthographic conventions therefore tend to regularize Garhwali forms toward Hindi norms, resulting in systematic deletion or insertion of the halant. These error patterns are consistent with phonological and orthographic differences between Hindi and Garhwali.

6 Conclusion

In this work, we present a comprehensive study of dialectal speech recognition across a diverse set of Devanagari-script Indic language varieties, with a particular focus on understanding how dialectal variation interacts with cross-lingual transfer in low-resource ASR. Through a large-scale empirical evaluation, we find that while ASR performance is generally associated with phylogenetic distance across languages, this factor alone does not explain performance in the dialectal setting. In particular, when evaluating on dialects in the zero-shot setting, we observe lower word error rates when the fine-tuning language is a dialect or nonstandard variety. In many cases, fine-tuning on small amounts of dialectal speech yields performance comparable to or better than fine-tuning on larger amounts of phylogenetically closer, high-resource standardized languages.

Across multiple phylogenetic subgroups, our results consistently demonstrate that including higher-resource mainstream languages during fine-tuning does not reliably improve zero-shot ASR performance on dialectal evaluation sets. Instead, whether the fine-tuning data itself reflects dialectal speech emerges as a more informative predictor of performance than phylogenetic proximity alone. These findings highlight the importance of treating dialects as distinct acoustic and linguistic entities rather than as minor variants of standardized languages when designing ASR systems.

We further present the first detailed ASR analysis for Garhwali, a nonstandard Pahari language variety, and show that a `w2vBERT`-based model fine-tuned on Garhwali achieves the best performance among the evaluated architectures. Although the resulting word error rate of 49.3% remains insufficient for fully automated transcription, this case study illustrates both the challenges of dialectal ASR and the benefits of dialect-specific modeling. Our quantitative error analysis further reveals substantial bias toward Hindi in both multilingual and Hindi-fine-tuned models, manifest-

ing in systematic normalization of dialectal and code-mixed forms, and underscoring the need for dialect-aware data selection and modeling strategies in future ASR systems.

Overall, our findings suggest that effective ASR for low-resource dialects requires moving beyond default assumptions of phylogenetic similarity and toward evaluation and modeling practices that explicitly account for dialectal variation.

Limitations

Our study is based on the VAANI dataset, which contains varying amounts of data across language varieties. Although this diversity allows us to evaluate ASR performance across a wide range of realistic dialectal settings, differences in dataset size may influence performance comparisons across languages. We mitigate this effect where possible by controlling the amount of fine-tuning data used across languages, though some variability remains inherent to the dataset.

In addition, VAANI consists of spontaneous and naturally-occurring speech collected across diverse regions. While this enables evaluation under realistic acoustic and conversational conditions, the presence of background noise, disfluencies, and region-specific recording environments may introduce additional variability in model performance.

Our analysis is limited to Indic dialects and language varieties written in the Devanagari script. Although this choice allows for controlled comparisons within a shared orthographic system, it excludes Indic languages written in other scripts, and our findings may not directly generalize beyond the Devanagari-script subset.

Finally, our experiments focus on a specific set of self-supervised ASR architectures and fine-tuning strategies. While these models are representative of widely used contemporary approaches, different architectures or training objectives may exhibit different transfer behaviors.

References

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 *Proc. Interspeech*, pages 2278–2282.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Proc. NeurIPS*, volume 33, pages 12449–12460.

Verena Blaschke, Miriam Winkler, and Barbara Plank. 2025. Standard-to-dialect transfer trends differ across text and speech: A case study on intent and topic classification in German dialects. *Preprint*, arXiv:2510.07890.

Yu-An Chung, Yu Zhang, Wei Han, Chung-Cheng Chiu, James Qin, Ruoming Pang, and Yonghui Wu. 2021. w2v-BERT: Combining contrastive learning and masked language modeling for self-supervised speech pre-training. In *Proc. IEEE ASRU*, pages 244–250.

Anuj Diwan, Rakesh Vaideeswaran, Sanket Shah, Ankita Singh, Srinivasa Raghavan, Shreya Khare, Vinit Unni, Saurabh Vyas, Akash Rajpuria, Chiranjeevi Yarra, Ashish Mittal, Prasanta Kumar Ghosh, Preethi Jyothi, Kalika Bali, Vivek Seshadri, Sunayana Sitaram, Samarth Bharadwaj, Jai Nanavati, Raoul Nanavati, and Karthik Sankaranarayanan. 2021. MUCS 2021: Multilingual and code-switching asr challenges for low resource Indian languages. In *Proc. Interspeech*, pages 2446–2450.

Rachana Gusain, Satya Ranjan Dash, Shantipriya Parida, and Girish Nath Jha. 2023. Automatic language identification: a case study of Pahari languages. *Language Resources and Evaluation*, 57:1361–1387.

Harald Hammarström, Robert Forkel, Martin Haspelmath, and Sebastian Bank. 2024. Indo-aryan. Glottolog 5.0.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. HuBERT: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:34513460.

Anjali Kantharuban, Ivan Vulić, and Anna Korhonen. 2023. Quantifying the dialect gap and its correlates across languages. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Lachman M. Khubchandani. 1991. India as a sociolinguistic area. *Language Sciences*, 13(2):265–288.

Gokul Karthik Kumar, V PraveenS., Pratyush Kumar, Mitesh M. Khapra, and Karthik Nandakumar. 2022. Towards building text-to-speech systems for the next billion users. In *Proc. ICASSP*, pages 1–5.

Saurabh Kumar, Amartyaveer, and Prasanta Kumar Ghosh. 2025. Jointly Improving Dialect Identification and ASR in Indian Languages using Multimodal Feature Fusion. In *Proc. Interspeech*, pages 2770–2774.

Colin Massica. 1993. *The Indo-Aryan Languages*. Cambridge University Press.

Iuliia Nigmatulina, Tannon Kew, and Tanja Samardzic. 2020. ASR for non-standardised languages with dialectal variation: the case of Swiss German. In *Proc. VarDial*, pages 15–24.

Hemant Palivela, Meera Narvekar, David Asirvatham, Shashi Bhushan, Vinay Rishiwal, and Udit Agarwal. 2025. Code-switching ASR for Low-Resource Indic Languages: A Hindi-Marathi case study. *IEEE Access*, 13:9171–9198.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *Proc. ICML*.

Mk Riyal, Vinod Khanduri, Nikhil Rajput, and Nagma Irfan. 2016. Creation and analysis of (agriculturally) speech database for Uttarakhand. *Indian Journal of Industrial and Applied Mathematics*, 7:181.

Harveen Singh Chadha, Anirudh Gupta, Priyanshi Shah, Neeraj Chhimwal, Ankur Dhuriya, Rishabh Gaur, and Vivek Raghavan. 2022. Vakyansh: ASR Toolkit for Low Resource Indic languages. *arXiv e-prints*, arXiv:2203.16512.

Chihiro Taguchi and David Chiang. 2024. Language complexity and speech recognition accuracy: Orthographic complexity hurts, phonological complexity doesn't. In *Proc. ACL*, pages 15493–15503, Bangkok, Thailand.

VAANI. 2025. VAANI: Capturing the language landscape for an inclusive digital India (phase 1). https://vaani.iisc.ac.in/.

A Appendix

A.1 Baseline Evaluation

We used the `IndicWav2Vec-Hindi` model to establish a baseline for the performance of off-the-shelf models on the languages in the VAANI dataset. We further divided the results into two tables based on whether the Evaluation Language was seen by the `IndicWav2Vec-Hindi` model during pretraining or not.

Language	CER	WER
Bhojpuri	0.786	0.732
Chhattisgarhi	0.783	0.757
Hindi	0.752	0.504
Konkani	0.851	0.916
Maithili	0.763	0.696
Marathi	0.821	0.870
Nepali	0.785	0.957
Rajasthani	0.768	0.762

Table 7: IndicWav2Vec-Hindi performance on languages seen during pre-training.

Language	CER	WER	Language	CER	WER
Angika	0.791	0.725	Khortha	0.829	0.743
Awadhi	0.774	0.771	Kumaoni	0.759	0.596
Bajjika	0.784	0.784	Kurukh	0.825	0.875
Bhili	0.860	0.944	Magadhi	0.782	0.700
Bundeli	0.774	0.646	Malvani	0.808	0.881
Garhwali	0.760	0.820	Marwari	0.752	0.740
Gondi	0.718	0.945	Nagpuri	0.797	0.834
Halbi	0.788	0.873	Sadri	0.742	0.738
Haryanvi	0.811	0.766	Surgujia	0.762	0.747
Jaipuri	0.696	0.532	Surjapuri	0.829	0.880
Khariboli	0.784	0.605	Thethi	0.796	0.707

Table 8: IndicWav2Vec-Hindi performance on related but unseen languages.

A.2 Orthographic Analysis

We computed the total number of words (tokens), total number of unique words (types) and the total number of words that only appear once (Hapax) per all the languages in the VAANI dataset. We also compute the total percentage of words that occur only once (Hapax%) and the ratio of unique words to total number of words (TTR) in the table 9.

A.3 Garhwali ASR

We evaluated multiple model architecture varieties on the Garhwali language test split from the VAANI dataset. For each model architecture, we first evaluated the performance without finetuning first and then after finetuning on the Training split of the Garhwali dataset. For each finetuning experiment except for the `Whisper` model, we generated the vocab of characters present in the training and validation split.

Language	Tokens	Types	Hapax	Hapax %	TTR
Gondi	3,136	1,562	1,168	74.78	0.4981
Thethi	1,556	654	469	71.71	0.4203
Bhili	1,420	632	448	70.89	0.4451
Kurukh	2,022	798	559	70.05	0.3947
Nepali	2,169	776	536	69.07	0.3578
Surjapuri	2,450	860	593	68.95	0.3510
Sadri	6,599	2,096	1,417	67.60	0.3176
Nagpuri	1,571	612	404	66.01	0.3896
Malvani	9,123	2,501	1,636	65.41	0.2741
Konkani	31,668	6,327	4,055	64.09	0.1998
Awadhi	2,346	887	561	63.25	0.3781
Surgujia	5,611	1,524	950	62.34	0.2716
Bundeli	7,734	1,613	969	60.07	0.2086
Khariboli	15,345	2,465	1,476	59.88	0.1606
Angika	34,348	5,420	3,233	59.65	0.1578
Garhwali	77,030	10,431	6,216	59.59	0.1354
Haryanvi	3,747	983	584	59.41	0.2623
Halbi	16,716	3,351	1,979	59.06	0.2005
Marathi	171,424	16,876	9,919	58.78	0.0984
Bajjika	32,049	4,681	2,717	58.04	0.1461
Magadhi	55,891	7,101	4,093	57.64	0.1271
Khortha	38,324	4,499	2,584	57.43	0.1174
Jaipuri	2,074	599	339	56.59	0.2888
Chhattisgarhi	169,861	13,024	7,287	55.95	0.0767
Bhojpuri	209,523	15,243	8,508	55.82	0.0728
Marwari	85,883	8,279	4,582	55.34	0.0964
Kumaoni	22,911	3,190	1,737	54.45	0.1392
Maithili	194,541	14,317	7,779	54.33	0.0736
Hindi	156,389	8,156	4,381	53.72	0.0522
Rajasthani	116,180	7,988	4,273	53.49	0.0688

Table 9: Hapax-Legomena (unique words) per training set.

Model Type	Tr. Lang	FT Lang	Gen Vocab?	Metric	Tr. Error	Val Loss	Tr. Loss	Steps	Test WER	Test CER
wav2vec2CTC	-	Garhwali	Yes	WER	0.769	1.576	0.403	6500	0.769	-
wav2vec2CTC	-	-	Yes	-	-	-	-	-	1.015	2.436
XLS-R	-	Garhwali	Yes	WER	0.735	1.554	0.171	1600	0.650	0.270
XLS-R	-	-	Yes	-	-	-	-	-	1.000	1.292
wav2vec2 BERT	-	Garhwali	Yes	CER	0.197	0.940	0.397	1200	**0.493**	**0.193**
wav2vec2 BERT	-	-	Yes	-	-	-	-	-	1.000	1.600
Whisper-small	Hindi	Garhwali	No	CER	0.242	0.974	0.001	4000	0.629	0.650
Whisper-small	Hindi	-	No	-	-	-	-	-	2.518	1.262
HuBERT	-	Garhwali	Yes	CER	-	-	-	-	0.515	0.199

Table 10: Comparison of Speech Models (Garhwali Fine-tuning vs. Baselines)