Whispering in Ol Chiki: Cross-Lingual Transfer Learning for Santali Speech Recognition

Atanu Mandal[1], Madhusudan Ghosh[2], Pratick Maiti*[1, 3], Sudip Kumar Naskar[1]

[1]Jadavpur University, Kolkata, INDIA,

[2]Indian Association for the Cultivation of Science, Kolkata, INDIA,

[3]Ramakrishna Mission Vidyamandira, Howrah, INDIA

Correspondence: atanumandal0491@gmail.com

Abstract

India, a country with a large population, possesses two official and twenty-two scheduled languages, making it the most linguistically diverse nation. Despite being one of the scheduled languages, Santali remains a low-resource language. Although Ol Chiki is recognized as the official script for Santali, many continue to use Bengali, Devanagari, Odia, and Roman scripts. In tribute to the upcoming centennial anniversary of the Ol Chiki script, we present an Automatic Speech Recognition for Santali in the Ol Chiki script. Our approach involves cross-lingual transfer learning by utilizing the Whisper framework pre-trained in Bengali and Hindi on the Santali language, using Ol Chiki script transcriptions. With the adoption of the Bengali pre-trained framework, we achieved a Word Error Rate (WER) score of 28.47 %, whereas the adaptation of the Hindi pre-trained framework resulted in a score of 34.50 % WER. These outcomes were obtained using the Whisper Small framework.

1 Introduction

Speech recognition has emerged as an important technology in the field of human-computer interaction, bridging the gap between spoken language and digital systems. With the advent of advanced deep learning, Automatic Speech Recognition (ASR) systems have been significantly improved, achieving human-level performance for widely spoken languages such as English, Mandarin, and Spanish (Graves et al., 2013; Amodei et al., 2016; Baevski et al., 2020). However, developing robust ASR systems for low-resource languages remains a challenging task due to the scarcity of annotated datasets, linguistic resources, and pre-trained language models (Besacier et al., 2014; Arivazhagan et al., 2019). One such low-resourced language is Santali, which is predominantly spoken by approximately 7.6 million people in India, Bangladesh, Nepal, and Bhutan. Despite its recognition as one of India's important languages, technological advancements in speech processing for Santali are still in an early stage.

Existing research in speech recognition for low-resource languages have explored various modeling techniques, including Hidden Markov Models (HMM) (Rabiner, 1989), Gaussian Mixture Models (GMM) (Reynolds et al., 2009), and deep learning based frameworks such as Transformers and Convolutional Neural Networks (CNN) (Graves et al., 2006; Gulati et al., 2020). For instance, Singh et al. (2023) demonstrated the efficacy of model adaptation for Bengali and Bhojpuri, while Priya et al. (2022) improved ASR performance using sequence modelling and transformer-based spell correctors. Javed et al. (2024a) created the "LAHAJA" benchmark dataset to enable evaluation of Hindi ASR systems on multiple accents. The dataset contains read and extempore speech on a diverse set of topics and use cases. Additionally, Shetty and Sagaya Mary N.J. (2020) highlighted the advantages of multilingual frameworks for low-resource Indian languages. Joshi et al. (2025) introduced "SRUTI", a benchmark comprising speech from rural Bhojpuri women speakers, aimed at facilitating voice-based access to agriculture, finance, government schemes, and healthcare services. Kumar et al. (2026) developed an ASR system for Sanskrit, addressing the distinctive challenges posed by the language's intricate linguistic and morphological characteristics. Kumar et al. (2022) used subword tokenisation strategies and search space enrichment with linguistic information to address the challenges posed by high degree of out-of-vocabulary entries in Sanskrit.

Existing studies on Santali have focused on language processing tools, such as a finite-state mor-

*The work was carried out while the author was at Ramakrishna Mission Vidyamandira. The author is currently affiliated with Jadavpur University.

Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, pages 269–278
December 20-24, 2025 ©2025 Association for Computational Linguistics

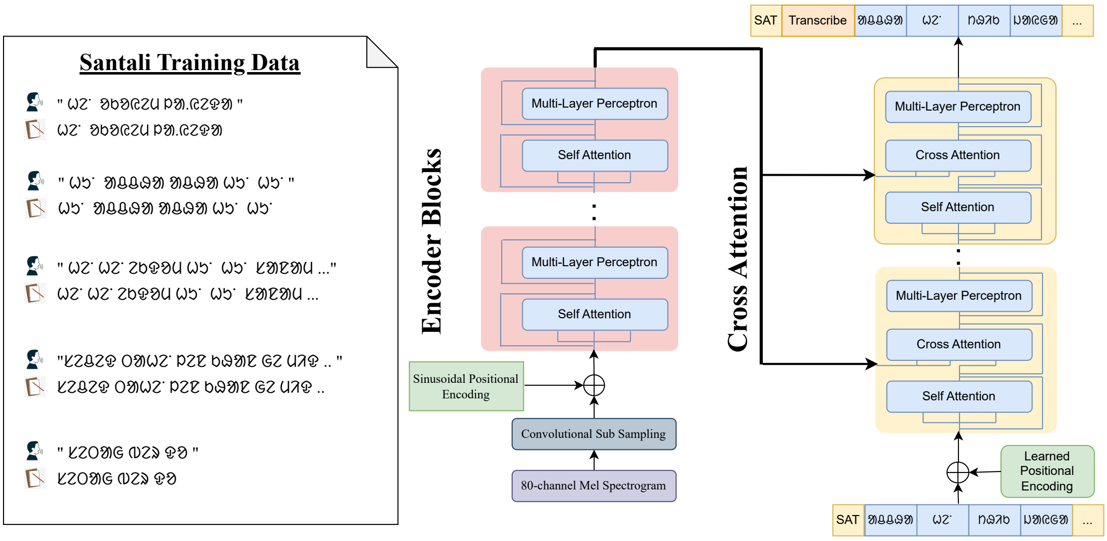

Figure 1: Overview of the Whisper-based ASR system fine-tuned for Santali speech recognition. The input audio is converted into an 80-channel Mel spectrogram and processed by convolutional sub-sampling and sinusoidal positional encoding. The encoder, composed of Transformer blocks with self-attention and multi-layer perceptrons, extracts audio features. The decoder, with self-attention, cross-attention, and learned positional encoding, generates character-level transcriptions in the Ol Chiki script, guided by cross-attention between encoder and decoder representations.

phological analyzer by Akhtar et al. (2017) and a dialect classifier using deep autoencoders by Sahoo et al. (2021). In ASR, Kumar et al. (2020) showed that triphone models outperform monophone models for Santali digits in Roman script. However, despite these advancements, the development of ASR systems specifically developed for Santali remains largely unexplored. Existing approaches have either relied on Roman or regional scripts such as Bengali, Hindi, and Odia, neglecting the Ol Chiki script of Santali.

Our investigations distinguish themselves by focusing on Santali speech transcribed in the Ol Chiki script, unlike previous studies that used the Roman script, bridging a crucial gap in ASR research. Our approach addresses these limitations by fine-tuning OpenAI's Whisper framework (Radford et al., 2022), a state-of-the-art (SOTA) ASR model. We used pre-trained in Bengali and Hindi, two linguistically and geographically proximate languages, to enhance the recognition of Santali phonetic patterns, applying cross-lingual transfer learning to improve ASR performance. Unlike previous works, we leverage Whisper's multilingual capabilities to adapt the model for Santali ASR in the Ol Chiki script. This approach marks a significant step toward developing inclusive and accurate speech recognition systems for the Santali-speaking community, addressing both linguistic diversity and technological accessibility. Our work not only advances the field of low-resource ASR but also sets a precedent for future research on indigenous languages, ensuring that linguistic diversity is preserved and celebrated in the digital age.

Our Contributions: The primary contributions of our work are summarized as follows:

- We develop the first ASR system specifically for Santali speech in the Ol Chiki script, marking a significant step toward digital inclusion for the Santali-speaking community.
- Our approach employs cross-lingual transfer learning by fine-tuning Whisper models pretrained in Bengali and Hindi, achieving WERs of 28.47% and 34.50%, respectively, demonstrating the effectiveness of linguistic proximity in low-resource scenarios.
- We provide a comprehensive evaluation of various Whisper model sizes (Tiny, Base, Small, Medium, Large), mentioning the trade-offs between model complexity and recognition performance.
- We studied the impact of LoRA-based parameter efficient fine-tuning on various Whisper models (Tiny, Base, Small, Medium, Large).

2 Language Perspective

The official script for the Santali language is Ol Chiki. Pandit Raghunath Murmu proposed the script in 1925. The shapes of the Ol Chiki characters are believed to be inspired by nature, physical forms, and the daily life of the Santals. The same principle applies to the sounds represented by these symbols. For example, the pronounced sound /at/ (ᱛ) is depicted by a circle, whose shape symbolizes the earth, and the meaning of the sound matches this representation. Likewise, the letter /ut/ (ᱩ) resembles the shape and sound of a mushroom. Ol Chiki is written from left to right and consists of six vowels and twenty-four consonants with five basic diacritics. The letters are arranged in a 6 by 5 matrix, where the first letter of each row, or the first column, represents the vowels, while the remaining letters are consonants. Furthermore, three vowels can be formed using the diacritic /gahla tudag/ (ᱹ), which can follow the vowels /la/ (ᱚ), /laa/ (ᱟ), and /le/ (ᱮ). The diacritic /mu tudag/ (ᱸ) nasalized vowels, and the combination of /mu tudag/ (ᱸ) and /gahla tudag/ (ᱹ) create a nasalized version of a newly formed vowel. The other three diacritics—/rela/ (ᱻ), /phaarkaa/ (ᱼ), and /ahad/ (ᱽ) —serve as a length marker, glottal protector, and deglottalization, respectively. Ol Chiki also includes two punctuation marks, /mucaad/ (᱾) and double /mucaad/ (᱿), both used in poetry, while only /mucaad/ (᱾) is employed in prose to indicate the end of a sentence. Latin punctuation marks such as commas, question marks, exclamation marks, parentheses, and quotation marks are also utilized. Lastly, Ol Chiki employs the decimal number system and has its own set of numerals (᱐, ᱑, ᱒, ᱓, ᱔, ᱕, ᱖, ᱗, ᱘, ᱙).

Despite belonging to a different language family, the prolonged interaction between Santali speakers and those of Indo-Aryan languages such as Bengali, Odia, and Hindi has led to some similarities in speech. However, Santali retains its uniqueness in fundamental linguistic structure, grammar, and vocabulary. Here are some key areas of similarity in speech:

1. Sentence Structure and Syntax
   - Subject-Object-Verb (SOV) Order
     Like Bengali, Odia, and Hindi, Santali follows the SOV word order. The sentences in all four languages typically place the subject first, followed by the object, and the verb at the end.

2. Pronunciation and Accent
   - Consonant Sounds:
     Santali exhibits patterns of aspirated and unaspirated consonants comparable to those found in Bengali, Odia, and Hindi. The aspirated sounds (like /p^{h}/ (ᱯᱷ), /b^{h}/ (ᱵᱷ), /k^{h}/ (ᱠᱷ)) in these languages contribute to a similar pronunciation style, particularly in formal or deliberate speech. Additionally, the retroflex consonants characteristic of Hindi and Odia are also present in Santali.
   - Nasalization:
     Santali displays a significant use of nasalized sounds, a feature also found in Bengali and, to a lesser extent, Odia. This nasalization influences the pronunciation of vowels, imparting a melodic quality comparable to that of the spoken forms of these languages. Although Hindi has fewer nasalized vowels compared to Santali and Bengali, nasalization does occur in specific contexts.

3. Intonation and Rhythm
   - Melodic Patterns:
     Santali and Bengali, in particular, possess a melodious and flowing intonation that gives the spoken languages a softer and more rhythmic quality. Odia exhibits a similar trait in informal conversation, whereas Hindi tends to be more monotonic and straightforward. Although the tonal quality of Santali speech is not as pronounced as in tonal languages, it has been shaped by the influence of neighboring languages, especially Bengali.
   - Stress and Lengthening of Syllables:
     The inclination to elongate specific syllables in both Santali and Bengali contributes to a rhythmic quality in their speech. For example, vowel lengthening is a prevalent feature in spoken Bengali and Santali, where vowels are extended in certain contexts for emphasis or to adhere to the phonological rules of the language. Although Odia exhibits some of this trait, Hindi generally features less vowel elongation.

4. Code-Switching and Borrowed Vocabulary
   - Shared Loanwords
     As a result of significant interaction between the Santali-speaking community and speakers of Bengali, Odia, and Hindi, Santali has adopted numerous words from these

languages, particularly for contemporary concepts, administration, and technology. In urban or bilingual settings, speakers frequently code-switch between Santali and the neighboring Indo-Aryan languages.

3 Methodology

Task Description: The objective of this study is to develop an ASR system tailored specifically for the Santali language in the Ol Chiki script. Given an audio input sequence $X = \{x_1, x_2, \ldots, x_T\}$, $x_t \in \mathbb{R}^d$, where T is the number of time steps and d is the feature dimension, the system aims to predict the corresponding text transcription. The goal is to generate a sequence of characters $Y = \{y_1, y_2, \ldots, y_L\}$, $y_l \in \mathcal{V}$, where L is the number of characters and $\mathcal{V}$ denotes the vocabulary of Ol Chiki characters. The ASR model aims to maximize the conditional probability $P(Y \mid X; \theta) = \prod_{l=1}^{L} P(y_l \mid X, y_1, \ldots, y_{l-1}; \theta)$, where θ denotes the model parameters.

3.1 Encoder-Decoder Framework

Our proposed ASR system is built upon the Whisper (Radford et al., 2022) framework, which is an encoder-decoder model. The overview of our framework is shown in Figure 1. The model is fine-tuned on Santali speech data using cross-lingual transfer learning from pre-trained Bengali and Hindi models due to their proximity and phonetic similarities.

Feature Extraction: The audio waveform is first preprocessed to standardize the input features. Each audio sample is resampled to a sampling rate of 16 kHz and converted to a 16-bit mono channel. Then, an 80-channel log-Mel spectrogram, $X \in \mathbb{R}^{T \times 80}$, is computed for the input to the encoder.

Encoder: The encoder processes the input spectrogram using N Transformer blocks. Each block consists of a multi-head self-attention layer and a feedforward neural network with residual connections:

$$H_0 = X,$$

$$H_n = \text{LayerNorm}\big(H_{n-1} + \text{SelfAttention}(H_{n-1})\big)$$

$$H_n = \text{LayerNorm}\big(H_n + \text{FFN}(H_n)\big), n = 1, \ldots, N$$

where $\text{SelfAttention}(H)$ is computed as:

$$\text{SelfAttention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right) V$$

with query Q, key K, and value V matrices obtained from the input H.

Decoder: The decoder autoregressively generates text output one token at a time by applying masked multi-head attention. Given the encoded representation H_N, the decoder generates output tokens as:

$$Z_0 = \text{Embedding}(y_{\text{<start>}})$$

$$Z_l = \text{LayerNorm}(Z_{l-1} + \text{MaskedAttention}(Z_{l-1}))$$

$$Z_l = \text{LayerNorm}(Z_l + \text{CrossAttention}(Z_l, H_N)),$$

$$l = 1, \ldots, L$$

where $\text{CrossAttention}(Z, H_N)$ is defined as:

$$\text{CrossAttention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right) V$$

Finally, a linear layer followed by a softmax function is applied to predict the next character:

$$P(y_l \mid X, y_1, \ldots, y_{l-1}) = \text{softmax}(W_o Z_l + b_o)$$

Training Procedure: The model is fine-tuned using the cross-entropy loss function:

$$\mathcal{L} = -\sum_{l=1}^{L} \log P(y_l \mid X, y_1, \ldots, y_{l-1})$$

The final layer of the pre-trained Whisper Small model is fine-tuned while all other layers are frozen.

Inference: During inference, the decoder generates tokens sequentially using greedy decoding:

$$\hat{y}_l = \arg\max_{y_l \in \mathcal{V}} P(y_l \mid X, \hat{y}_1, \ldots, \hat{y}_{l-1})$$

4 Experiment Set Up

4.1 Dataset Description

For experimental validation, we used the Santali Speech Dataset with the Ol Chiki script transcriptions, compiled from two publicly accessible sources: Mozilla Common Voice[1] (Ardila et al., 2020) and AI4Bharat IndicVoices (Javed et al., 2024b). On average, Common Voice training segments are 4.8 seconds long (~6 words), while IndicVoices training segments are longer at 6.2 seconds (~12 words). For evaluation, we used the Common Voice test set, which spans 5.3 seconds (~6 words) on average. Dataset statistics for training, validation, and test splits are provided in Table 1.

[1] The latest Common Voice dataset was extracted on July 03, 2025, from Link.

Table 1: Summary of the Santali speech corpus used for training and evaluation. The table lists the number of audio samples in the training, validation, and test sets. Note that the test set for IndicVoices is not yet released ([a]).

Sl. No.	Corpus Name	Train	Valid	Test
1	IndicVoices (Javed et al., 2024b)	19,779	249	-[a]
2	Common Voice (Ardila et al., 2020)	333	68	127
	Total	20,112	317	127

4.2 Implementation Details

Table 2: Architecture parameter(s) of the Whisper framework

Framework	No. of Layers	Width	No. of Heads	Parameters
Tiny	4	384	6	39M
Base	6	512	8	74M
Small	12	768	12	244M
Medium	24	1024	16	769M
Large	32	1280	20	1550M

The training parameters of the Whisper framework are summarized in Table 2. Fine-tuning was performed using a learning rate of 1×10^{-5} with the “AdamW” optimizer. Training was conducted for 40 epochs with a batch size of 16. Only the final layer was updated during training, while all the other layers were frozen. Since Santali is not among the supported languages in Whisper, we used models pre-trained in Bengali and, for comparison, fine-tuned a model pre-trained in Hindi on Santali data. We fine-tuned pre-trained ASR models in Hindi and Bengali, as these languages are linguistically and geographically close to Santali. The Bengali and Hindi pretraining refers to the internal representation already available in Whisper for these languages, not separate fine-tuned checkpoints. Our codes are available in the following link[2].

4.3 Research Questions

To systematically investigate the effectiveness of cross-lingual transfer learning for ASR in the Santali language using the Ol Chiki script, we formulate the following research questions. These questions aim to analyze the impact of source language proximity, model architecture size, dataset characteristics, and fine-tuning strategies on the overall performance of the adapted Whisper models.

- **RQ1:** Which language, Bengali or Hindi, provides better cross-lingual transfer learning performance for Santali speech recognition, and what factors contribute to this difference?
- **RQ2:** How does the model size (Tiny, Base, Small, Medium, Large) influence the WER when fine-tuned with Bengali and Hindi pre-trained models, and why does the Small variant outperform the others?
- **RQ3:** How do different datasets (Common Voice vs. IndicVoices) affect the fine-tuning performance of the Whisper model, and what dataset characteristics contribute to the observed WER differences?
- **RQ4:** How does Parameter Efficient Fine-Tuning (PEFT), specifically LoRA fine-tuning, perform in low-resource dataset scenarios, and what factors contribute to the observed results?

5 Results

In this section, we provide all the findings of the experiments. For evaluation purposes, we used the Common Voice Test, which contains 127 samples.

Table 3: WER (in %) of different Whisper model variants without fine-tuning, using Bengali and Hindi pre-trained checkpoints on the Santali speech dataset. This table provides baseline performance across model sizes before any task-specific adaptation.

Framework	Bengali pre-Trained without Fine-Tuning	Hindi pre-Trained without Fine-Tuning
Tiny	201.12	201.12
Base	197.05	197.05
Small	111.64	111.64
Medium	115.99	115.99
Large	**108.42**	**108.42**

Table 3 provides the evaluation results of Whisper frameworks done on the Bengali pre-trained and Hindi pre-trained models. The results show that the increase in parameter sizes decreases the WER but yet is unable to recognise the required transcriptions. This is due to the non-presence of the Santali language in Whisper-trained languages.

Language Comparison: Bengali vs. Hindi (RQ1): In response to **RQ1**, Tables 4 and 5 show that the Bengali pre-trained Whisper Small model achieves a lower WER (28.47%) compared to the

[2] https://github.com/atanumandal0491/Santali-ASR

Hindi pre-trained model (34.50%) on the Common Voice Training Dataset. This performance gap is due to the greater phonetic and syntactic similarity between Bengali and Santali, such as shared vowel nasalization, consonant structures, and SOV word order, which facilitate more effective model adaptation during fine-tuning. Similarly, using the IndicVoices Training Dataset, fine-tuned Whisper Base model for both Bengali pre-trained (54.28%) and Hindi pre-trained (53.30%) shows similar results. This is due to the increase in the robust Dataset sample, which provides a low-parameter model to optimise.

Table 4: Performance comparison (WER in %) of Whisper model variants fine-tuned on the Common Voice Santali corpus, using Bengali and Hindi pre-trained checkpoints. The table highlights model-wise effectiveness after full fine-tuning across both source languages.

Framework	Bengali pre-Trained with Full Fine-Tuning	Hindi pre-Trained with Full Fine-Tuning
Tiny	118.09	102.81
Base	101.54	98.04
Small	**28.47**	**34.50**
Medium	93.27	129.73
Large	32.96	35.34

Table 5: WER (in %) of trained IndicVoices Santali Corpus on Whisper Frameworks in the Bengali and Hindi pre-trained language.

Framework	Bengali pre-Trained with Full Fine-Tuning	Hindi pre-Trained with Full Fine-Tuning
Tiny	62.55	111.08
Base	**54.28**	**53.30**
Small	57.36	54.84
Medium	99.86	100.00
Large	117.67	112.06

Impact of Model Size (RQ2): For **RQ2**, Tables 4 and 5 show that the Bengali pre-trained Whisper Small model achieves the lowest WER—28.47% and Bengali pre-trained Whisper Small model 34.50%, outperforming both smaller (Tiny, Base) and larger (Medium, Large) variants. Its balanced architecture (12 layers, 768 hidden dimensions) allows it to capture phonetic patterns without overfitting effectively. In contrast, larger models are harder to optimize with limited data, while smaller ones lack sufficient capacity to model complex linguistic features. It also suggests that smaller models can capture complex linguistic feature if it is provided with robust large datasets.

Dataset Influence: Common Voice vs. IndicVoices (RQ3): For **RQ3**, Tables 4 and 5 show that fine-tuning on the Common Voice dataset yields lowers WERs (28.47% for the Bengali pre-trained Whisper model, 34.50% for the Hindi pre-trained Whisper model) than IndicVoices (54.28% and 53.30%, respectively). This performance gap is likely due to Common Voice's shorter utterances (4.8 seconds, ~6 words), which allow for more precise alignment between audio and text. In contrast, the longer and more variable utterances in IndicVoices (6.2 seconds, ~12 words) introduce complexity that challenges the model during training.

LoRA Finetuning (RQ4): To address RQ4, we evaluate the impact of LoRA-based parameter-efficient fine-tuning using Bengali and Hindi pre-trained Whisper models across different model sizes, as shown in Tables 6 and 7.

Table 6: Performance comparison (WER in %) of Whisper model variants fine-tuned using LoRA on the Common Voice Santali corpus, with Bengali and Hindi as source pre-trained languages. The table presents the impact of parameter-efficient fine-tuning across different model sizes.

Framework	Bengali pre-Trained with LoRA Fine-Tuning	Hindi pre-Trained with LoRA Fine-Tuning
Tiny	113.04	131.00
Base	185.41	158.35
Small	101.26	121.80
Medium	62.97	**98.60**
Large	**61.43**	99.58

Table 7: Performance comparison (WER in %) of Whisper model variants fine-tuned using LoRA on the IndicVoices Santali corpus, based on Bengali and Hindi pre-trained checkpoints. The table highlights model-wise adaptation under parameter-efficient fine-tuning in a low-resource setting.

Framework	Bengali pre-Trained with LoRA Fine-Tuning	Hindi pre-Trained with LoRA Fine-Tuning
Tiny	268.16	374.47
Base	134.92	108.56
Small	**121.18**	211.92
Medium	188.36	364.38
Large	126.51	**134.36**

A clear trend emerges where larger models (e.g., Medium and Large) consistently outperform smaller ones (Tiny, Base, Small) under LoRA fine-tuning, particularly on the Common Voice dataset. This can be attributed to the fact that larger models possess greater capacity to retain and adapt relevant linguistic patterns, even when only a subset of parameters specifically in the attention layers—is updated. However, despite this relative gain, none of the LoRA tuned models match the performance of their fully fine-tuned counterparts, underscoring LoRA's limited expressiveness when operating un-

Table 8: Performance comparison (WER in %) of the proposed Whisper-based models against state-of-the-art IndicConformer systems across Common Voice and IndicVoices datasets. The table summarizes results from both full fine-tuning and LoRA-based parameter-efficient tuning across different model sizes and pre-training languages.

Model Name	**Pretrained on**	**cf.**	**Dataset**	**WER**
IndicConformer-CTC[2]	–	–	IndicVoices	53.04
IndicConformer-RNNT[2]	–	–	IndicVoices	50.78
Whisper Full Finetune Small	Bengali	Table 4	Common Voice	**28.47**
Whisper Full Finetune Small	Hindi	Table 4	Common Voice	34.50
Whisper Full Finetune Base	Bengali	Table 5	IndicVoices	54.28
Whisper Full Finetune Base	Hindi	Table 5	IndicVoices	53.30
Whisper LoRA Finetune Large	Bengali	Table 6	Common Voice	61.43
Whisper LoRA Finetune Medium	Hindi	Table 6	Common Voice	98.60
Whisper LoRA Finetune Small	Bengali	Table 7	IndicVoices	121.18
Whisper LoRA Finetune Large	Hindi	Table 7	IndicVoices	134.36

der strict parameter constraints. The degradation is more noticeable in the IndicVoices dataset, where longer and acoustically varied utterances challenge the model's ability to generalize, especially when the fine-tuning signal is narrow. These results suggest that while LoRA offers an efficient and stable training paradigm suitable for large models in low-resource scenarios, it struggles to fully adapt to complex linguistic and phonetic variations without broader parameter updates.

For Benchmarking, we evaluated our results with the state-of-the-art IndicConformer[3] framework proposed by AI4Bharat. IndicConformer is a multilingual 130M conformer based model following the same architecture as proposed by Tjandra et al. (2023). Table 8 shows the benchmarking results.

6 Error Analysis

While metrics like Word Error Rate (WER) offer a broad view of model accuracy, they often miss the specific types of errors that impact usability. To address this, we analyzed individual outputs from the Bengali Common Voice evaluation set to better understand where the model performs well and where it breaks down. This section shows the Error Analysis of our best model, i.e., Whisper Small pre-trained in Bengali.

Table 9 shows examples of Common Errors the model made. This qualitative analysis revealed a number of patterns that highlight both the strengths and weaknesses of the system.

- Confusion Between Similar Sounding Characters

  In many cases, the model confused characters that sound alike, especially in fast or informal speech. For instance:

  Predictions : ᱥᱮᱱᱛᱟᱜ ᱠᱥᱚᱥᱩ ᱫᱟ ᱵᱤᱞᱣᱟᱮᱪᱟ ᱵᱟᱮᱟ।
  Reference : ᱥᱮᱱᱛᱟᱜ ᱠᱥᱚᱥᱩ ᱫᱟ ᱵᱤᱛᱞᱣᱟᱮᱪᱟ ᱵᱟᱮᱟ।
  WER: 20.0%

  Here, the model missed "ᱛ" before "ᱞ", likely due to phonetic similarity. Although the sentence is still mostly correct, this minor change subtly alters pronunciation and fluency.

- Errors in Suffixes and Grammatical Particles

  Bengali and Santali heavily rely on suffixes and particles to convey tense, mood, and case. The model often mishandled these, either by dropping, altering, or misplacing them.

  Predictions : ᱥᱮᱱᱛᱟᱜ ᱠᱥᱚᱥᱩ ᱫᱟ ᱵᱯᱟᱛᱤᱮᱟ ᱵᱟᱮᱟ।
  Reference : ᱥᱮᱱᱛᱟᱜ ᱠᱥᱚᱥᱩ ᱫᱟ ᱵᱯᱟᱱᱮᱤ ᱵᱟᱮᱟ।
  WER: 20.0%

  The substitution of "ᱱᱮᱤ" with "ᱛᱤᱮᱟ" suggests that the model has trouble preserving proper suffix morphology, especially in contexts where nasalization or tense is involved.

[3] Model is available at https://huggingface.co/ai4bharat/indicconformer_stt_sat_hybrid_ctc_rnnt_large.

Table 9: Example of prediction errors and their types.

Sl. No.	Reference Sentence	Predicted Sentence	Error Type	WER (%)
1	ᱩᱱᱤᱭᱟᱜ ᱦᱩᱪᱩᱢ ᱫᱚ ᱠᱟᱱᱩᱴᱷ ᱠᱟᱱᱟ᱾	ᱩᱱᱤᱭᱟᱜ ᱦᱩᱪᱩᱢ ᱫᱚ ᱠᱩᱱᱴᱟᱱ ᱠᱟᱱᱟ᱾	Consonant substitution ("ᱠᱟᱱᱩᱴᱷ" → "ᱠᱩᱱᱴᱟᱱ")	20.0
2	ᱩᱱᱤᱭᱟᱜ ᱦᱩᱪᱩᱢ ᱫᱚ ᱠᱷᱟᱭᱥᱤᱭᱟᱣ ᱠᱟᱱᱟ᱾	ᱩᱱᱤᱭᱟᱜ ᱦᱩᱪᱩᱢ ᱫᱚ ᱠᱟᱭᱥᱤᱭᱟ ᱠᱟᱱᱟ᱾	Suffix omission (missing "ᱣ")	20.0
3	ᱩᱱᱤᱭᱟᱜ ᱦᱩᱪᱩᱢ ᱫᱚ ᱠᱮᱜᱟᱱ ᱠᱟᱱᱟ᱾	ᱩᱱᱤᱭᱟᱜ ᱦᱩᱪᱩᱢ ᱫᱚ ᱠᱮᱨᱟᱱ ᱠᱟᱱᱟ᱾	Phonetic confusion ("ᱜ" → "ᱨ")	20.0
4	ᱟᱨ ᱵᱟᱵᱟ ᱡᱩᱱᱜᱮᱯ ᱪᱩᱥᱩᱭ ᱴᱟᱠᱟ ᱮᱢᱟ ᱫᱮᱭᱟᱭ᱾	ᱟᱨ ᱵᱟᱵᱟ ᱡᱩᱱᱜᱮᱯ ᱪᱩᱥᱩᱭ ᱴᱟᱠᱟ ᱮᱢᱟ ᱫᱮᱭᱟᱭ᱾	No error	0.0
5	ᱩᱱᱤᱭᱟᱜ ᱦᱩᱪᱩᱢ ᱫᱚ ᱠᱯᱤᱱᱮ ᱠᱟᱱᱟ᱾	ᱩᱱᱤᱭᱟᱜ ᱦᱩᱪᱩᱢ ᱫᱚ ᱠᱯᱤᱱᱮ ᱠᱟᱱᱟ᱾	No error	0.0

- Insertions and Omissions in Longer Sentences With longer sentences, the model occasionally skipped words or inserted unnecessary ones. These kinds of structural issues were more pronounced in complex phrases.

  Predictions : ᱵᱮᱜ ᱴᱷᱮᱜ ᱢᱤᱫᱛᱟᱭ ᱟᱥᱮ ᱜᱮᱯ ᱡᱮ ᱴᱟᱠᱟ ᱜᱮ ᱛᱟᱥᱮᱨ ᱠᱟᱱᱟ᱾
  Reference : ᱵᱮᱠ ᱴᱷᱮᱱ ᱢᱤᱫᱛᱟᱭ ᱟᱥᱮ ᱜᱮᱯ ᱡᱮ ᱴᱟᱠᱟ ᱜᱮ ᱛᱟᱥᱮᱨ ᱠᱟᱱᱟ᱾
  WER: 20.0%

  The replacement of "ᱵᱮᱠ" with "ᱵᱮᱜ" and "ᱴᱷᱮᱱ" with "ᱴᱷᱮᱜ" changed the meaning of the sentence and introduced fluency issues. Such mistakes indicate that the model may have difficulty aligning longer sequences during decoding.

- Difficulty with Morphologically Complex Words In morphologically rich contexts, particularly those involving compounding or inflexions, the model's performance dropped. This is a known challenge in low-resource settings and is reflected in errors like this:

  Predictions : ᱩᱱᱤᱭᱟᱜ ᱦᱩᱪᱩᱢ ᱫᱚ ᱠᱮᱵᱥᱤᱱ ᱠᱟᱱᱟ᱾
  Reference : ᱩᱱᱤᱭᱟᱜ ᱦᱩᱪᱩᱢ ᱫᱚ ᱠᱮᱵᱥᱭᱤᱱ ᱠᱟᱱᱟ᱾
  WER: 20.0%

  Here, the model omits the "ᱭ" character in "ᱠᱮᱵᱥᱭᱤᱱ", possibly simplifying the form. However, in doing so, it loses grammatical correctness.

From these examples, we can see that the model generally performs well on shorter, simpler sentences, but its accuracy declines when handling:

- Phonetic similarities that lead to substitutions
- Morphological variations, particularly in suffixes and particles
- Longer utterances in which insertions and omissions become more common

These issues highlight the challenges of working with morphologically rich and phonologically complex languages, such as Santali, especially under low-resource conditions.

7 Conclusions & Future Work

In this paper, we present an initial but important effort in developing an ASR system for Santali using the Ol Chiki script. By fine-tuning the Whisper framework with cross-lingual transfer learning on Bengali and Hindi, we have demonstrated the feasibility of creating accurate speech recognition models for under-resourced languages. Our findings indicate that fine-tuning the Whisper Small model on the Common Voice dataset yields the most promising results, achieving WERs of 28.47% and 34.50% with Bengali and Hindi pretraining, respectively. These results demonstrate that transfer learning offers a viable path to address the ASR challenges faced by under-resourced languages, significantly improving access to digital technologies for their speakers while preserving linguistic diversity.

Although this study provides a strong foundation for Santali ASR, several areas can be explored for future research. These include:

- **Expanding Training Data.** The performance of the ASR system could be further improved by increasing the size and diversity of the Santali speech dataset.

- **Exploring Other Pre-trained Models.** While this work focused on Bengali and Hindi pre-trained models, exploring other linguistically related languages could potentially yield better results.

- **Adapting the Model for Different Accents and Dialects.** Santali exhibits regional variations in pronunciation and vocabulary. Future research could focus on adapting the ASR system to better handle these variations through techniques such as transfer learning or domain adaptation.

- **Incorporating a Language Model.** Integrating a language model trained on Santali text data could help improve the accuracy of the ASR system by providing contextual information and reducing word error rates.

By addressing these challenges and pursuing future research directions, we can further advance the Santali ASR field and contribute to preserving and promoting this valuable language.

Limitations

Our study makes a meaningful contribution to speech technology for the Santali language, but it has certain limitations. These include

- The scope of our experiments is constrained by the limited size and diversity of available Santali speech data, particularly in the "Ol Chiki" script. This limitation may impact the generalisation of the model to broader dialectal and acoustic variations within the Santali-speaking population.

- Although our approach leverages cross-lingual transfer from Bengali and Hindi due to their linguistic proximity to Santali, these source languages are not perfectly aligned regarding phonetic and syntactic characteristics. As a result, some Santali-specific nuances may not be fully captured by the adapted models.

- The evaluation is limited to the Whisper Small variant. Although we briefly explored models of varying sizes, comprehensive tuning and optimization of larger or alternative architectures were outside the scope of this work due to computational constraints.

Acknowledgments

We are deeply grateful to the Santali-speaking community, whose language and culture continue to inspire this work. This research is especially dedicated to the Ol Chiki script, created by Pandit Raghunath Murmu in 1925. As *the centenary of the Ol Chiki script is being celebrated in 2025*, our work aims to make a meaningful contribution to the growing movement to bring Santali onto the digital map and celebrate its unique linguistic identity.

This research was partially funded by the 'VIDYAAPATI: Bidirectional Machine Translation Involving Bengali, Konkani, Maithili, Marathi, and Hindi' under the Project titled 'National Language Translation Mission (NLTM): BHASHINI'.

References

Md. Amir Khusru Akhtar, Mohit Kumar, and Gadadhar Sahoo. 2017. Automata for santali language processing. In *2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, pages 939–943.

Dario Amodei, Sundaram Ananthanarayanan, Rishita Anubhai, Jingliang Bai, Eric Battenberg, Carl Case, Jared Casper, Bryan Catanzaro, Qiang Cheng, Guoliang Chen, and 1 others. 2016. Deep speech 2: End-to-end speech recognition in english and mandarin. In *International conference on machine learning*, pages 173–182. PMLR.

R. Ardila, M. Branson, K. Davis, M. Henretty, M. Kohler, J. Meyer, R. Morais, L. Saunders, F. M. Tyers, and G. Weber. 2020. Common voice: A massively-multilingual speech corpus. In *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, pages 4211–4215.

Naveen Arivazhagan, Ankur Bapna, Orhan Firat, Dmitry Lepikhin, Melvin Johnson, Maxim Krikun, Mia Xu Chen, Yuan Cao, George Foster, Colin Cherry, and 1 others. 2019. Massively multilingual neural machine translation in the wild: Findings and challenges. *arXiv preprint arXiv:1907.05019*.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460.

Laurent Besacier, Etienne Barnard, Alexey Karpov, and Tanja Schultz. 2014. Automatic speech recognition for under-resourced languages: A survey. *Speech communication*, 56:85–100.

Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. 2006. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd international conference on Machine learning*, pages 369–376.

Alex Graves, Abdel-rahman Mohamed, and Geoffrey Hinton. 2013. Speech recognition with deep recurrent neural networks. In *2013 IEEE international conference on acoustics, speech and signal processing*, pages 6645–6649. Ieee.

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

Tahir Javed, Janki Nawale, Sakshi Joshi, Eldho George, Kaushal Bhogale, Deovrat Mehendale, and Mitesh M. Khapra. 2024a. LAHAJA: A Robust Multi-accent Benchmark for Evaluating Hindi ASR Systems. In *Interspeech 2024*, pages 2320–2324.

Tahir Javed, Janki Atul Nawale, Eldho Ittan George, Sakshi Joshi, Kaushal Santosh Bhogale, Deovrat Mehendale, Ishvinder Virender Sethi, Aparna Ananthanarayanan, Hafsah Faquih, Pratiti Palit, Sneha Ravishankar, Saranya Sukumaran, Tripura Panchagnula, Sunjay Murali, Kunal Sharad Gandhi, Ambujavalli R, Manickam K M, C Venkata Vaijayanthi, Krishnan Srinivasa Raghavan Karunganni, and 2 others. 2024b. Indicvoices: Towards building an inclusive multilingual speech dataset for indian languages. *Preprint*, arXiv:2403.01926.

Sakshi Joshi, Eldho Ittan George, Tahir Javed, Kaushal Bhogale, Nikhil Narasimhan, and Mitesh M. Khapra. 2025. Recognizing Every Voice: Towards Inclusive ASR for Rural Bhojpuri Women. In *Interspeech 2025*, pages 4243–4247.

Arvind Kumar, Rampravesh Kumar, and Kamlesh Kishore. 2020. Performance analysis of asr model for santhali language on kaldi and matlab toolkit. In *2020 International Conference on Recent Trends on Electronics, Information, Communication & Technology (RTEICT)*, pages 88–92.

Rishabh Kumar, Devaraja Adiga, Rishav Ranjan, Amrith Krishna, Ganesh Ramakrishnan, Pawan Goyal, and Preethi Jyothi. 2022. Linguistically Informed Post-processing for ASR Error correction in Sanskrit. In *Interspeech 2022*, pages 2293–2297.

Rishabh Kumar, Devaraja Adiga, Rishav Ranjan, Amrith Krishna, Ganesh Ramakrishnan, Pawan Goyal, and Preethi Jyothi. 2026. Linguistically informed automatic speech recognition in sanskrit. *Computer Speech & Language*, 95:101861.

M.C.S. Priya, D.K. Renuka, and L.A. Kumar. 2022. Multilingual low resource indian language speech recognition and spell correction using indic bert. *Sādhanā*, 47(4):227.

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

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision. *Preprint*, arXiv:2212.04356.

Douglas A Reynolds and 1 others. 2009. Gaussian mixture models. *Encyclopedia of biometrics*, 741(659-663).

Sunil Kumar Sahoo, Brojo Kishore Mishra, Shantipriya Parida, Satya Ranjan Dash, Jatindra Nath Besra, and Esaú Villatoro Tello. 2021. Automatic dialect detection for low resource santali language. In *2021 19th OITS International Conference on Information Technology (OCIT)*, pages 234–238.

Vishwas M. Shetty and Metilda Sagaya Mary N.J. 2020. Improving the performance of transformer based low resource speech recognition for indian languages. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8279–8283.

Abhayjeet Singh, Arjun Singh Mehta, Ashish Khuraishi K S, Deekshitha G, Gauri Date, Jai Nanavati, Jesuraja Bandekar, Karnalius Basumatary, Karthika P, Sandhya Badiger, Sathvik Udupa, Saurabh Kumar, Savitha, Prasanta Kumar Ghosh, Prashanthi V, Priyanka Pai, Raoul Nanavati, Rohan Saxena, Sai Praneeth Reddy Mora, and Srinivasa Raghavan. 2023. Model adaptation for asr in low-resource indian languages. *Preprint*, arXiv:2307.07948.

Andros Tjandra, Nayan Singhal, David Zhang, Ozlem Kalinli, Abdelrahman Mohamed, Duc Le, and Michael L. Seltzer. 2023. Massively multilingual asr on 70 languages: Tokenization, architecture, and generalization capabilities. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.