Cross-Lingual Transfer Learning for Santali Speech Recognition

Anonymous ACL submission

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 23.59 %, whereas the adaptation of the Hindi pre-trained framework resulted in a score of 28.75 % WER. These outcomes were obtained using the Whisper Small framework.

1 Introduction

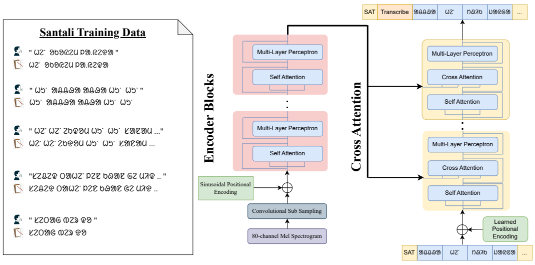

Figure 1: Overview of the Whisper-based ASR system fine-tuned for Santali speech recognition. The input audio is converted into an 80-channel Mel spectrogram and processed by convolutional sub-sampling and sinusoidal positional encoding. The encoder, composed of Transformer blocks with self-attention 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.

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. Additionally, Shetty and Sagaya Mary N.J. (2020) highlighted the advantages of multilingual frameworks for low-resource Indian languages. Existing studies on Santali have focused on language processing tools, such as a finite-state morphological 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 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 for 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 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 pre-trained in Bengali and Hindi, achieving WERs of 23.59% and 28.75%, 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.

2 Proposed 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.

2.1 Encoder-Decoder Framework

Our proposed ASR system is built upon Whisper (Radford et al., 2022) framework, which is an encoder-decoder model. 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 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:

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

3 Experiment Set Up

3.1 Dataset Description

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., 2024)	45,389	485	-[a]
2	Common Voice (Ardila et al., 2020)	315	462	147
	Total	45,704	947	147

For experimental validation, we used the Santali Speech Dataset with the Ol Chiki script transcriptions, compiled from two sources: Mozilla Common Voice (Ardila et al., 2020) and AI4Bharat IndicVoices (Javed et al., 2024). On average, Common Voice segments last 4.3 seconds (6 words), while IndicVoices segments are longer at 6.4 seconds (13 words). Dataset statistics for training, validation, and test splits are provided in Table 1.

3.2 Research Questions

To explore the effectiveness of our finetuned ASR system for Santali using cross-lingual transfer learning, we propose the following research questions (RQs).

- **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 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?

3.3 Implementation Details

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. Only the final layer was updated during training, while all other layers were frozen. Since Santali is not among the supported languages in Whisper, we used models pre-trained in Bengali and, for comparison, also fine-tuned a Hindi pre-trained model on Santali data.

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

4 Results

In this section, each research question is discussed in detail, with key findings highlighted.

Table 3: WER (in %) of trained Santali Corpus on Whisper Small Framework in the Bengali pre-trained language.

No Fine-tuning	Common Voice Fine-tuning	IndicVoice Fine-tuning
189.93	**23.59**	44.28

Language Comparison: Bengali vs. Hindi (RQ1): In response to **RQ1**, Tables 3 and 6 show that the Whisper Small model fine-tuned with Bengali achieves a lower WER (23.59%) compared to the Hindi pre-trained model (28.75%). 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 facilitates more effective model adaptation during fine-tuning.

Table 4: WER (in %) of trained Santali Corpus on Whisper Framework (based on their sizes) in the Bengali pre-trained language.

Framework	No Fine-tuning	Fine-tuned
Tiny	124.57	97.42
Base	188.82	98.89
Small	189.93	**23.59**
Medium	211.18	99.51
Large	187.59	27.27

Table 5: WER (in %) of trained Santali Corpus on Whisper Framework (based on their sizes) in the Hindi pre-trained language.

Framework	No Fine-tuning	Fine-tuned
Tiny	124.57	101.11
Base	188.82	99.88
Small	182.05	**28.75**
Medium	190.05	100.00
Large	187.59	30.10

Table 6: WER (in %) of trained Santali Corpus on Whisper Small Framework in the Hindi pre-trained language.

No Fine-tuning	Common Voice Fine-tuning	IndicVoice Fine-tuning
182.05	**28.75**	45.67

Impact of Model Size (RQ2): For **RQ2**, Tables 4 and 5 show that the Whisper Small model achieves the lowest WER—23.59% for Bengali and 28.75% for Hindi, outperforming both smaller (Tiny, Base) and larger (Medium, Large) variants. Its balanced architecture (12 layers, 768 hidden dimensions) allows it to effectively capture phonetic patterns without overfitting. In contrast, larger models are harder to optimize with limited data, while smaller ones lack sufficient capacity to model complex linguistic features.

Dataset Influence: Common Voice vs. IndicVoices (RQ3): For **RQ3**, Tables 3 and 6 show that fine-tuning on the Common Voice dataset yields lower WERs (23.59% for Bengali, 28.75% for Hindi) than IndicVoices (44.28% and 45.67%, respectively). This performance gap is likely due to Common Voice's shorter utterances (4.3 seconds, 6 words), which allow for more precise alignment between audio and text. In contrast, the longer and more variable utterances in IndicVoices (6.4 seconds, 13 words) introduce complexity that challenges the model during training.

5 Conclusions & Future Work

This paper has presented 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 23.59% and 28.75% with Bengali and Hindi pre-training, 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 by preserving linguistic diversity. Although this study provides a strong foundation for Santali ASR, several areas are unexplored 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 these 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 it completes 100 years in 2025, we hope our work adds to the growing efforts to bring Santali into the digital age and celebrate its unique linguistic identity.

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 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. 2024. Indicvoices: Towards building an inclusive multilingual speech dataset for indian languages. *Preprint*, arXiv:2403.01926.

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.

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.