End-to-End Automatic Speech Recognition System for Sanskrit: Whisper-Based Modelling

[1]Devendr Kumar, [2]Girish Nath Jha, [3]Khalid Choukri
[1,2]School of Sanskrit and Indic Studies, Jawaharlal Nehru University, New Delhi, India
[3]Evaluations and Language Resources Distribution Agency (ELDA), 75013 Paris, France

devneed2@gmail.com , girishjha@gmail.com , choukri@elda.org

Abstract

Sanskrit, one of the world's oldest classical languages, continues to hold immense significance in religious, cultural, and academic traditions. Yet, despite this importance, it remains underrepresented in modern speech technologies due to the scarcity of large-scale datasets and ASR systems. In this paper, we present the development of an end-to-end Automatic Speech Recognition (ASR) system for Sanskrit using OpenAI's Whisper architecture. As part of this effort, we created and preprocessed a 212-hour aligned speech-text corpus, compiled from publicly available resources such as VākSancaya and Shrutilipi (Bhashini group), along with carefully curated manual recordings. The Whisper model was fine-tuned on this dataset and evaluated on an 11-hour test set comprising both prose and recitation. The system achieved a Word Error Rate (WER) of 9.43% and a Character Error Rate (CER) of 4.93%, demonstrating the effectiveness of Whisper for low-resource, morphologically complex languages. We further discuss the challenges encountered in Sanskrit data creation (phonetic coverage, sandhi phenomena, and speaker variation), model adaptation strategies, and potential applications in digital humanities, including oral text preservation and Sanskrit resource digitization.

Keywords: Sanskrit Automatic Speech Recognition, Whisper architecture, Low-resource languages, Oral text preservation

1. Introduction

Sanskrit, often regarded as the mother of many Indo-Aryan languages, has a rich oral and written tradition spanning millennia. It remains central to religious rituals, classical literature, philosophy, and linguistic studies. However, in the digital era, Sanskrit's representation in speech technologies has been limited. Unlike widely spoken languages such as English, Hindi, or Mandarin, Sanskrit suffers from a lack of large-scale annotated speech corpora and robust ASR systems. The development of Automatic Speech Recognition (ASR) systems for Sanskrit faces unique challenges: Phonetic richness with approximately 48 phonemes (Deshpande, 1975)., including aspirated, retroflex, and nasalized sounds. Morphological complexity with extensive use of inflection, compounding, and sandhi (euphonic combination). Variation in speaking styles, particularly between recitation, chanting, and prose reading.

To address these challenges, we explore the use of Whisper, a transformer-based ASR model by OpenAI, which has shown robustness in multilingual and noisy settings. This work contributes one of the first large-scale, end-to-end Sanskrit ASR systems trained on a 107-hour curated dataset, aiming to advance digital humanities and support applications such as oral text preservation, e-learning, and linguistic research.

2. Related Work

We utilized open-source datasets. The वाक्-सञ्चयः (/Vāksañcayaḥ/) dataset (Adiga, 2021) is an extensive resource for Sanskrit ASR. This dataset consists of 78 hours of speech data, covering around 46,000 sentences, making it a comprehensive dataset for ASR model training. It captures linguistic characteristics of Sanskrit, including phonemic orthography and Sandhi phenomena, ensuring that a wide range of speech patterns and linguistic features are represented. The dataset consists of recordings contributed by 27 speakers from six different native language backgrounds: Hindi, Kannada, Malayalam, Marathi,

Devendr Kumar, Girish Nath Jha, Khalid Choukri, *End-to-End Automatic Speech Recognition System for Sanskrit: Whisper-Based Modelling*. In: Z. Vetulani, P. Paroubek, I. Kardava (Eds.), Human Language Technologies as a Challenge for Computer Science and Linguistics – 2025, pp. 37–40. Poznań: Adam Mickiewicz University Press, 2025. © Authors, 2025. https://doi.org/10.14746/amup.9788323245414.7. Open Access chapter, distributed under the terms of the CC licence (BY-NC-ND, https://creativecommons.org/licenses/by-nc-nd/4.0/).

Tamil, and Telugu. These speakers include both native Sanskrit speakers and proficient non-native speakers, which tried to enhance the datasets' diversity and robustness for ASR training. This variety increases ability to generalize across different accents and speech patterns. Recordings were primarily conducted using Android phones via the Recorder app, the Audacity platform, and other online sources. Transcriptions were meticulously handled using tools like O Transcribe, ensuring high-quality data for training the model. This dataset provides a foundation for developing a robust Sanskrit ASR system, (See Table 1 for this details).

"IndicVoices" presents a comprehensive effort to create an inclusive multilingual speech dataset covering 22 Indian languages. This dataset, titled Indicvoices, (Javed, 2024) includes a total of 7348 hours of read, extempore, and conversational audio from 16,237 speakers across 145 Indian districts. Of these hours, 1639 have been transcribed, providing a rich resource for developing Automatic Speech Recognition (ASR) systems. The Sanskrit dataset was 27 hours of speech data within Indicvoices was developed with the help of Sanskrit Bharati. This specific dataset is a part of the broader effort to capture the linguistic diversity of India, including both high-resource and low-resource languages. The inclusion of Sanskrit is particularly significant given its historical and cultural importance, as well as the challenges associated with collecting high-quality spoken data for a language that is primarily used in liturgical and academic contexts.

3. Dataset Creation and Preprocessing

The JNU ASR Dataset is a meticulously curated corpus designed to advance research in Automatic Speech Recognition (ASR) for Sanskrit (Kumar, D., 2023). Comprising recordings of 33,543 text sentences, this dataset was collected in two distinct phases to ensure a diverse and representative sample. The participants were primarily Master's students of Sanskrit, aged 20 to 30, and native speakers of Hindi, Odia, Bengali, and Assamese, hailing from linguistically rich states such as Bihar, Uttar Pradesh, Haryana, Madhya Pradesh, Rajasthan, Odisha, Bengal, and Assam. This regional diversity enhances the dataset's robustness, making it an invaluable resource for training ASR systems in the context of linguistic and phonetic variability.

The first phase of data collection involved 128 participants, yielding 67 hours of audio. Each participant contributed at least 300 sentences, with text segments distributed consecutively across contributors. The recordings predominantly relied on mobile phones, resulting in hurdles such as varying audio formats, inconsistent technical expertise among participants, and challenges in data sharing. These issues necessitated extensive post-processing to create a clean and usable dataset. Despite these challenges, this phase laid a strong foundation for capturing natural speech nuances in Sanskrit, incorporating variations influenced by regional accents and linguistic backgrounds.

The second phase, involving 53 participants and generating 40 hours of audio, significantly streamlined the recording process. A custom-built web app audio recorder ensured consistency in file formats and optimized the overall pipeline. This innovation not only saved time but also mitigated the technical challenges encountered during the first phase. With standardized tools and procedures, this phase marked a leap forward in efficiency and data quality. The final JNU ASR Dataset, encompassing 107 hours of speech data, stands as a comprehensive and diverse resource, paving the way for developing robust ASR models for Sanskrit while addressing the linguistic and technical challenges unique to low-resource languages.

3.1 Corpus Volume

Different Sanskrit Speech Corpora for ASR

Name of the dataset	No. of Hours	No. of Speakers	No. of Sentences	Owner Of the Dataset
Vāksañcayah	78	27	46,000	IIT Bombay
IndicVoice	27	*	48069	Bhashini, MietY
JNU Data	107	181	33,543	JNU, New Delhi
Total	212			

Table 1. Speech Data

4. Methodology

Focus to the ground-level application of these theories, examining how these challenges manifest in actual

ASR development. The Whisper model has been selected for testing and experimentation to provide a hands-on evaluation of these linguistic complexities in practice. These metrics highlight the model's performance in addressing the linguistic challenges identified.

4.1 Whisper Model Architecture

Whisper is one of the groundbreaking models for automatic speech recognition (ASR). Whisper sets itself apart from previous models like Wav2Vec 2.0 in several significant ways. One of the most notable differences is in the data used for pre-training. While Wav2Vec 2.0 was trained on 60,000 hours of unlabelled audio data, whisper was trained on a staggering 680,000 hours of labelled audio-transcription data. This immense amount of data includes 117,000 hours of multilingual ASR data, allowing Whisper to recognize and transcribe speech in over 96 languages. This multilingual capability is particularly noteworthy as it extends support to many low-resource languages that often lack sufficient data for robust ASR performance. For this study, we fine-tuned the **Whisper-small** model (Gandhi, S.,2022) balancing performance and computational feasibility.

4.2 Training & Evaluation of Sanskrit ASR Model

We developed a Sanskrit ASR model by fine-tuning the Whisper-V3 pretrained model, which consists of a 32-layer architecture with approximately 1550 M parameters. The model was trained on a curated dataset of 212 hours of Sanskrit speech, partitioned using a 95:5 train–test split, resulting in 11 hours of held-out test data. We tuned the hyperparameters of the model as per our requirements. The impact of the model's hyperparameters on its performance is reflected in various key metrics such as loss, Word Error Rate (von Neumann, T., 2025) (WER), and Character Error Rate (CER). Fine-tuning the pre-trained models, i.e. Whisper v3 allowed us for the adaptation of the model to our specific datasets of Sanskrit, improving its performance in recognizing and transcribing Sanskrit speech.

It observed a rapid decrease in loss during the early stages of training on train set, followed by a more gradual decline. This behaviour is consistent with the high capacity of the model, supported by its deep architecture (32 layers each for the encoder and decoder) and high-dimensional feedforward networks. The initial sharp drop in loss suggests that the model quickly adapted to the new data, benefiting from the large number of attention heads and layers that allowed it to capture complex dependencies in the input.

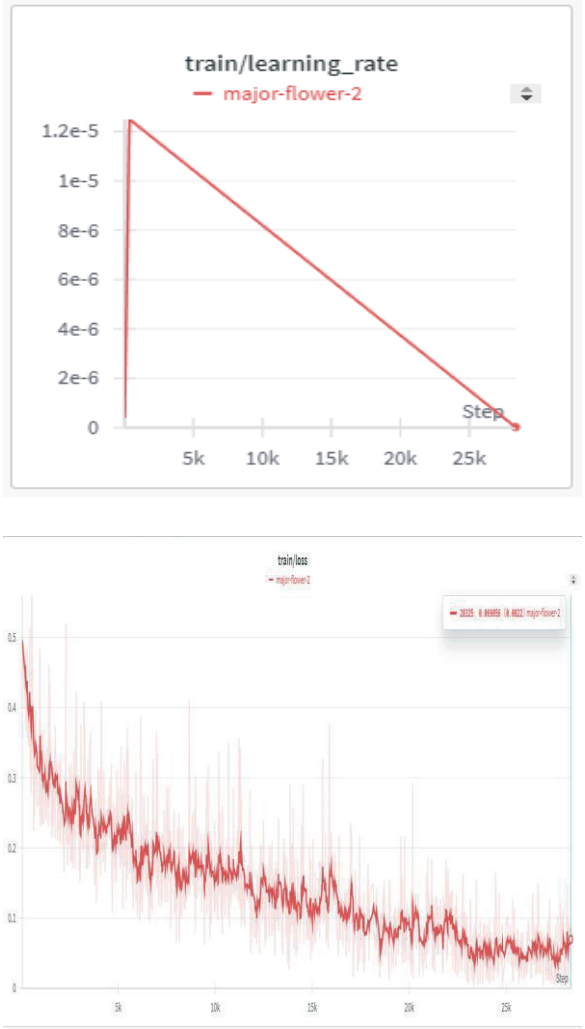

Fig.1: Training Loss

The model's performance was evaluated on unknown inputs by using the 11 hours of unseen data for automated evaluation. Following the application of the Whisper model, we observed a Word Error Rate (WER) of 9.43% and a Character Error Rate (CER) of 4.93%. Furthermore, we tested our user interface on six distinct speakers, allowing for a more realistic and

varied assessment of the system's performance in real-time situations.

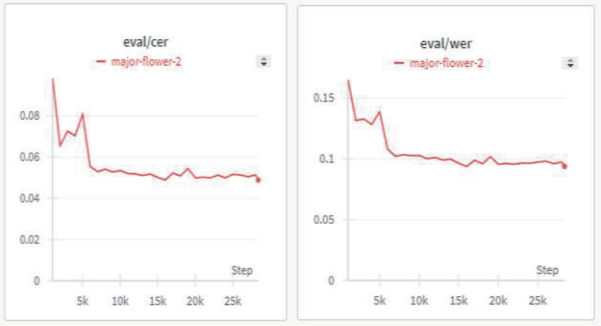

Fig. 2: Evaluation of WER and CER

4.3 Human Evaluation

Six distinct speakers' input was used to test our Sanskrit ASR system using Sanskrit User Interface. This method ensured a thorough review by enabling us to evaluate the ASR system's performance over a range of pronunciations, speech patterns, and speaking styles. Through human evaluation, the Sanskrit ASR system's performance was thoroughly examined. At the word level, Sanskrit speech recognition tends to achieve higher accuracy. However, when complete sentences are spoken, speakers often use compound words (samāsika padas), which increases the likelihood of recognition errors. Furthermore, pronunciation in continuous speech is not always phonetically precise, as speakers come from varied linguistic backgrounds and geographical regions throughout India. As a result, error rates are higher at the sentence level, while word-level recognition remains comparatively more accurate. The assessment showed a high degree of transcribing accuracy with a WER of 7.58% and a CER of 5.30%. Significant advancements in ASR development for Sanskrit are reflected in these measurements, which show the system's strong performance in correctly identifying and transcribing Sanskrit speech.

5. Conclusion

This paper presented one of the first large-scale end-to-end Sanskrit ASR systems using Whisper. With a **2**12-hour curated corpus**,** the system achieved WER 9.43% and CER 4.93%**,** demonstrating the feasibility of adapting Whisper to low-resource, morphologically complex languages. The work lays the foundation for future efforts in Sanskrit speech technology, contributing to digital humanities, education, and preservation of oral traditions.

6. References

Deshpande, M. (1975). Phonetics of short A in Sanskrit. *Indo-Iranian Journal*, *17*(3), 195-209. https://www.jstor.org/stable/24652891?seq=1

Adiga, D., Kumar, R., Krishna, A., Jyothi, P., Ramakrishnan, G., & Goyal, P. (2021). Automatic speech recognition in Sanskrit: A new speech corpus and modelling insights. *http://www.jstor.org/stable/24652891*

Adiga, D., Kumar, R., Krishna, A., Jyothi, P., Ramakrishnan, G., & Goyal, P. (2021). Automatic speech recognition in Sanskrit: A new speech corpus and modelling insights. *https://arxiv.org/abs/2106.05852* .

Javed, T., Nawale, J. A., George, E. I., Joshi, S., Bhogale, K. S., Mehendale, D., ... & Khapra, M. M. (2024). Indicvoices: Towards building an inclusive multilingual speech dataset for indian languages. *https://arxiv.org/abs/2403.01926* .

Kumar, D. Jha, G.N., (2023). Resource Creation for Sanskrit ASR (Automatic Speech Recognition). *An International Journal of Engineering Science. http://ijoes.vidyapublications.com/paper/Vol35/10.pdf* .

Gandhi, S., Von Platen, P., & Rush, A. M. (2022). Esb: A benchmark for multi-domain end-to-end speech recognition. *https://arxiv.org/abs/2210.13352* .

von Neumann, T., Boeddeker, C., Delcroix, M., & Haeb-Umbach, R. (2025). Word Error Rate Definitions and Algorithms for Long-Form Multi-talker Speech Recognition. *IEEE Transactions on Audio, Speech and Language Processing*. *https://arxiv.org/abs/2508.02112*