Development of Speech Corpus for Low-Resource Language- A case of Sanskrit

Devendr Kumar[1], Girish Nath Jha[2], Khalid Choukri[3]

Department of Language Technology and Language Engineering, Mahatma Gandhi Antarrashtriya Hindi Vishwavidyalaya, Wardha, Maharashtra[1]
School of Sanskrit and Indic Studies, Jawaharlal Nehru University, New Delhi[2]
European Language Resource Association, Paris[3]
devneed2@gmail.com[1], girishjha@gmail.com[2], choukri@elda.org[3]

Abstract

This paper presents a comprehensive framework for the development of a speech corpus for Sanskrit, designed to facilitate advances in Automatic Speech Recognition (ASR) and AI/ML research. The proposed corpus comprises over 107 hours of transcribed speech data, collected from diverse Sanskrit sources through a systematic and scalable pipeline. We detail the end-to-end methodology adopted for corpus creation, encompassing web crawling, data sanitization, audio downloading, and transcription alignment. Particular emphasis is placed on the methodological rigor applied at each stage, including source selection, preprocessing for quality assurance, transcription protocols, and forced alignment techniques. The paper further addresses the unique complexities inherent to Sanskrit, spanning its phonetic richness, intricate morphological structure, and distinctive syntactic patterns. By systematically addressing these dimensions, the resulting 107-hour corpus aims to serve as a foundational resource for speech technology research in Sanskrit.

Keywords: Sanskrit Speech Corpus, Resource creation for Sanskrit, Sanskrit speech recognition, Sanskrit Data, Sanskrit Text, aligned data.

1. Introduction

Sanskrit, one of the oldest and most grammatically sophisticated languages of the Indo-Aryan family, occupies a position of profound cultural, religious, and linguistic significance in the Indian subcontinent. Recognized as one of the 22 scheduled languages under the Eighth Schedule of the Indian Constitution, Sanskrit has served for millennia as the medium of philosophy, science, literature, and religious thought. Its grammatical framework, codified by the legendary grammarian Pāṇini in the 7th century BCE, remains unparalleled in its precision, drawing admiration from modern linguists and computer scientists alike.

Despite its rich legacy, Sanskrit today faces the paradox of reverence without widespread spoken use. Its everyday spoken application has declined significantly over the centuries, rendering it predominantly a written and liturgical language. This limited oral presence has hindered the development of modern speech technologies for Sanskrit, making the compilation of large-scale spoken language data a persistent challenge.

With the rapid advancement of Natural Language Processing (NLP) and AI/ML-driven speech technologies, languages such as English, Mandarin, and Hindi now benefit from robust speech corpora powering virtual assistants, transcription tools, and language learning platforms.

Proceedings of WILDRE-8, pages 55–60
12 May 2026. ©ELRA Language Resources Association (ELRA), 2026

However, Sanskrit remains largely underrepresented in this digital revolution. To address this gap, the present work describes the systematic development of a 107-hour Sanskrit speech corpus, compiled from diverse online sources through a pipeline involving web crawling, data sanitization, audio processing, transcription, and forced alignment. The corpus aims to facilitate advances in Automatic Speech Recognition (ASR), Text-To-Speech (TTS) synthesis, and broader linguistic research, while contributing to the digital preservation of the Sanskrit language.

The remainder of this paper is organized as follows: Section 2 reviews related work, Section 3 describes the data collection methodology, Section 4 details the preprocessing and alignment pipeline, Section 5 presents corpus statistics, Section 6 discusses challenges, and Section 7 concludes with future directions.

2. Related Work

There have been efforts to create Sanskrit speech corpora, and some related work has been done. The **AI4Bharat (**Javed, et. al. 2024) program has been started at IIT Madras for creating language resources for various Indian languages, including Sanskrit. They have worked on building a vertical corpus for Sanskrit, which includes both text and speech data. Sanskrit and other 11 Indian languages are represented in the **Shrutilipi** (Chadha, et. al. 2022) tagged ASR dataset which was created by mining parallel audio and text pairings at the document scale from All India Radio news bulletins. 27 hours (Gupta, et. al. 2021) of tagged speech data is available for Sanskrit. The **Vāksañcayaḥ** (Adiga, et. al. 2021) (Holla, et. al. 2022). speech corpus contains 45,953 sentences that were recorded at a sampling rate of 22 KHz over the course of 78 hours of data. Readings from diverse texts from Sanskrit literature make up the majority of the content.

3. Text Selection

Sanskrit encompasses a diverse literary tradition, spanning a vast range of genres, styles, and historical periods. At the broadest level, Sanskrit literature is categorized into two major forms: Vedic and Classical.

Vedic Sanskrit represents the earliest stratum of the language and comprises the four foundational Vedas- the Rigveda, Yajurveda, Samaveda, and Atharvaveda along with associated texts such as the Brahmanas, Aranyakas, and Upanishads. This body of literature is primarily composed of hymns, ritual prescriptions, and philosophical discourses, and stands as one of the oldest recorded literary traditions in human history.

Classical Sanskrit, on the other hand, is broadly divided into two principal literary forms: Prose and Poetry. For the purpose of the present corpus, Prose Sanskrit also commonly referred to as spoken Sanskrit was selected as the primary source of data. This decision was informed by the fact that Sanskrit Automatic Speech Recognition (ASR) research is still in its nascent stages, and spoken or prose-form text offers a more practical and representative foundation for building speech datasets.

In terms of data sourcing, openly available texts (Kumar, et. al. 2023) free from copyright restrictions were prioritized. These include content from Sanskrit Wikipedia, which provides a substantial collection of freely usable linguistic material. Additionally, a broad range of public domain literature was explored, encompassing classical texts, historical works, and other writings available across various digital repositories. Throughout this process, careful consideration was given to the ethical implications of collecting and

utilizing textual data, ensuring that all sourced material adheres to applicable legal and ethical standards.

3.1 Text Collection Tools

For large-scale text collection, the present study employed the IL Crawler[1] tool a specialized utility designed to systematically gather, scrape, and extract textual data from a broad spectrum of digital sources. These sources include websites, online documents, social media platforms, and various other repositories accessible on the internet. The tool operates by accepting URLs as input and autonomously retrieving and compiling the relevant linguistic material from the specified locations.

The primary strength of the IL Crawler lies in its ability to automate the otherwise labour-intensive and time-consuming process of data acquisition. By eliminating the need for manual data collection, the tool ensures a high degree of efficiency, consistency, and scalability in corpus construction qualities that are particularly critical when building large-scale linguistic resources such as a Sanskrit speech corpus.

4. Audio Recording

We obtained audio recordings of the collected text from a total of 181 speakers. In the initial stage, 112 speakers participated. The second stage involved 53 speakers. In the final stage, 16 speakers were engaged. The development of high-quality audio recordings requires meticulous attention to both the equipment used and the recording environment. For this purpose, we carefully selected appropriate equipment and optimized the recording conditions. The choice of microphone is particularly critical in determining recording quality. We employed clip-on microphones in conjunction with closed-back headphones, allowing for hands-free recording, real-time audio monitoring, and effective isolation from external noise.

Equally important was the selection of a recording location with minimal background interference. To ensure clarity, we avoided areas affected by traffic, electrical appliances, or other disruptive noise sources. Furthermore, recording rooms were acoustically optimized by closing windows and doors and applying weather stripping to seal gaps, thereby reducing external disturbances.

Through this rigorous setup, we ensured that the recordings achieved the necessary standard of clarity and consistency, providing a reliable foundation for subsequent speech processing tasks.

4.1 Speaker Selection and Variation

Careful attention was given to speaker selection in order to ensure a diverse and representative corpus. Speakers were recruited from a wide range of age groups, spanning from 15 to 45 years, to capture variation in voice characteristics, speech patterns, and pronunciation styles across different life stages. To promote regional diversity, participants were selected from multiple states and regions across India, representing a broad spectrum of geographical and cultural backgrounds. Furthermore, speakers with varying native language backgrounds were included, reflecting India's rich multilingual landscape and accounting for the influence of different mother tongues on Sanskrit pronunciation and intonation.

In addition to regional and linguistic diversity, gender balance was maintained throughout the selection process. An equal

[1] https://sanskrit.jnu.ac.in/download/index.jsp (Accessed on 12/04/2026)

representation of male and female speakers was ensured across all phases of recording, enabling the corpus to capture the full range of acoustic and phonetic variation associated with gender. This carefully balanced and inclusive approach to speaker selection significantly enhances the robustness, diversity, and generalizability of the corpus, making it a more reliable resource for training and evaluating Sanskrit speech recognition and synthesis systems.

4.2 Audio Recorder

To streamline the audio recording process, a dedicated web application recorder was developed using a Python Django backend server. This solution was specifically designed to address several key challenges inherent in large-scale speech data collection, including displaying text to participants, capturing audio recordings, standardizing file naming conventions, maintaining quality standards, and efficiently uploading and organizing the recorded files. The system automatically generates metadata including participant IDs and sentence IDs immediately after each recording is completed, enabling real-time tracking and organization of collected data. The recorded audio files are stored in a structured directory hierarchy, with a parent folder containing individual subfolders for each participant, ensuring clean and systematic file management.

In the context of the present corpus, this web application served as the primary recording platform for both Phase 2 and Phase 3 of data collection. The tool ensured consistent audio quality and uniform file formats across all recordings, significantly reducing the time and effort required for post-processing. The automated file naming and metadata generation further eliminated manual errors, improved data organization, and enabled faster file retrieval. By centralizing and standardizing the entire recording workflow, the web application substantially enhanced the efficiency, reliability, and scalability of the corpus development process.

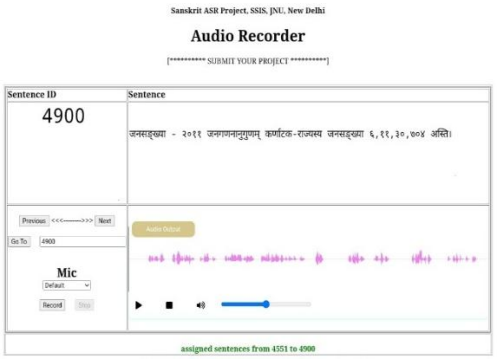

Figure 1: The Audio Recorder

4.3 Audio Editing and Formatting

The process of manipulating audio recordings plays a crucial role in improving quality, enhancing content, and preparing the data for linguistic applications. For this purpose, we employed the audio editing software Audacity along with an online audio converter, selected on the basis of functionality, reliability, and user familiarity. Essential editing tasks included adjusting the volume, panning, and timing of each track to achieve the desired sound balance and clarity. All recorded audio files were standardized by converting them into the .WAV format, a widely accepted format in automatic speech recognition research. Furthermore, parameters such as sample rate, bit depth, and encoding specifications were carefully configured in accordance with the requirements of speech processing systems. To facilitate efficient organization and categorization, comprehensive metadata was appended to each file, including details such as the recorder's name, gender, age, and region of origin. This metadata significantly enhances the dataset's utility for linguistic and sociolinguistic analysis.

Following the editing process, each audio file was subjected to critical listening to ensure that it satisfied the established

standards of clarity and accuracy. Additionally, recordings were tested across multiple playback devices to verify consistency and quality under different listening conditions. This rigorous post-processing pipeline ensured that the final dataset was both technically robust and linguistically valuable.

Phase	No. of speakers	No. of audio files	Duration (hours)
1	112	20,763	67
2	53	9,828	30
3	16	2,952	10
Total	**181**	**33,543**	**107**

Table 1: overview of JNU Data Set

4.4 Size of Speech Corpus

The size of a speech corpus is generally measured in terms of the total duration of audio recordings it contains. In the present work, a total of 33,543 audio files were recorded from 181 speakers, resulting in more than 107 hours of original in-house speech data. In addition, two publicly available Sanskrit speech corpora were collected from online sources, contributing a further 105 hours of data. Altogether, the present corpus comprises approximately 212 hours of Sanskrit speech data, making it one of the largest Sanskrit speech resources developed to date.

5. Conclusion

This paper presents the development of a comprehensive Sanskrit speech corpus aimed at advancing computational processing and digital preservation of Sanskrit. The corpus comprises a total of 212 hours of spoken Sanskrit data — of which 107 hours were recorded in-house from 181 speakers across three systematic phases at JNU, New Delhi, while the remaining 105 hours were sourced from publicly available online corpora. Together, these resources constitute one of the largest Sanskrit speech datasets developed to date.

The corpus development pipeline encompassing web crawling, text sanitization, audio recording, transcription, and forced alignment was designed with rigor and scalability in mind. The introduction of a custom-built web application recorder in Phases 2 and 3 standardized the recording process, reduced costs significantly, and improved overall data quality and consistency. The resulting corpus holds considerable promise for a wide range of applications including Automatic Speech Recognition (ASR), Text-To-Speech (TTS) synthesis, linguistic research, and digital preservation of Sanskrit manuscripts. Beyond its technological utility, this work carries deeper cultural significance helping bridge the gap between an ancient language and modern speech technology.

6. Future Work

The present corpus opens several promising directions for future research. Most immediately, the data will be used to train and evaluate Sanskrit ASR models, teaching systems to accurately recognize and transcribe spoken Sanskrit. The corpus will equally support the development of Sanskrit-capable voice assistants, enabling natural language understanding and response in Sanskrit. Beyond speech technology, linguists and sociolinguists can leverage this resource to study regional dialects, accents, and language variations among Sanskrit speakers. Expanding the corpus in terms of speaker diversity and domain coverage remains a long-term goal,

with the aim of building an increasingly robust foundation for Sanskrit speech technology research.

7. References

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. *arXiv preprint arXiv:2403.01926*.

Chadha, H. S., Gupta, A., Shah, P., Chhimwal, N., Dhuriya, A., Gaur, R., & Raghavan, V. (2022). Vakyansh: ASR Toolkit for Low Resource Indic languages. *arXiv preprint arXiv:2203.16512*.

Gupta, A., Chadha, H. S., Shah, P., Chhimwal, N., Dhuriya, A., Gaur, R., & Raghavan, V. (2021). Clsril-23: Cross lingual speech representations for indic languages. *arXiv preprint arXiv:2107.07402*.

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. *arXiv preprint arXiv:2106.05852*.

Holla, S. S., Kumar, T. M., Hiretanad, J. R., Deepak, K. T., & Narasimhadhan, A. V. (2022, March). End-to-end speech recognition for low resource language sanskrit using self-supervised learning. In *2022 International Conference on Wireless Communications Signal Processing and Networking (WiSPNET)* (pp. 148-152). IEEE.

Kumar, D. & Jha G. N. (2023). Resource Creation for Sanskrit ASR (Automatic Speech Recognition). *An International Journal of Engineering Science,* 35, 103-109.