Supplementary Materials for "Automatic Speech Recognition in Sanskrit: A New Speech Corpus and Modelling Insights"

Anonymous ACL-IJCNLP submission

1 Distribution of word length in three ASR datasets

Char count(N)	Sanskrit	Telugu	Gujarati
N<=6	32.92%	29.86%	47.86%
6<N<=12	47.31%	59.04%	47.64%
N>12	19.77%	11.1%	4.5%

Table 1: Distribution of Number of characters (*wrt* SLP1) per word in three ASR datasets

1.1 Differences between Sanskrit and other Indic languages for ASR

Many Indian languages are known to be derived from Sanskrit (Kulkarni et al., 2010) and their scripts derived from the Brahmi script (Salomon, 1996; Sproat, 2003), which leads to grapheme-based similarites amongst them. In Figure 1, we illustrate through an example, the spectrum of mapping the native character/grapheme (units) in words across languages; at one end of the spectrum is राम(/rām/) in Hindi mapped to రాము(/rāma/) in Telugu as an example where direct correspondence with the native character exists. Going further in the spectrum are examples for which direct character correspondence does not exist. सीता(/sītā/) in Hindi going to ಸೀತೆ(/sīte/) in Kannada is an instance where there is a change in the ending vowel.

Schwa Deletion The schwa deletion phenomenon plays a crucial role in the north Indian languages. Every consonant by itself includes a short /a/ vowel sound (referred to as "schwa") unless otherwise specified. For example, the letter 'त' in Hindi is pronounced as /ta/. This sound can be associated with any

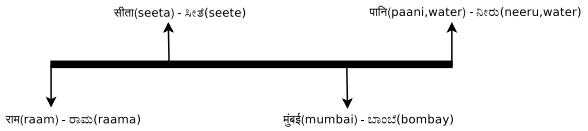

Figure 1: Spectrum of mapping native character/grapheme (units) in words across Indian languages

other vowel sound by the use of "Mātras". Mātras are dependent forms of vowels. Schwa is the default vowel for a consonant and hence does not require any explicit Mātra to represent it. Schwa deletion is a phenomenon where implicit schwas of a word are deleted during pronunciation. For example, in Hindi, the proper noun, 'अर्जुन (/arjun/, the name of a person) has schwa deletion after the consonant 'न' and is pronounced as Arjun. This phenomenon is not observed in the South Indian languages. For instance, in Kannada it is pronounced as 'Arjuna'. There is no implicit schwa deletion in Sanskrit as well as in the traditional use of South Indian languages such as Kannada. North Indian languages observe schwa deletion not only at the end of the word, but also in the middle of a word in some cases. For example, the word 'गलती' (/galtī/ meaning mistake) in Hindi observes implicit schwa deletion after the consonant 'ल' (/la/).

ASR becomes challenging because of this phenomenon since the occurrence of schwa deletion is not always explicitly specified in the orthography. For example, the name रामबाबु (/rāmbābu/) has two basic words concatenated to form a name. In Hindi, this name has an implicit schwa deleted at म (consonant sounding 'ma') of राम (/rām/). While constructing phonetic representations for ASR, such deletions introduce ambiguities in pro-

nunciation which could be alleviated by enforcing more consistency between graphemes and phonemes. This same word रामबाबु written in Telugu would be phonetically represented as రామ్బాబు (/rāmbābu/) instead of రామబాబు (/rāmabābu/) which is intuitive. Note that in the former case, there is an addition of ‘్’(*halant*: an explicit schwa deletion marker) at మ(/ma/). This forces the consonants మ(/ma/) and బ(/ba/) to combine and form a conjunct. In the latter case there is a grapheme consistency across both Hindi and Telugu languages but there is a variation in their pronunciation due to the schwa deletion phenomenon. In contrast, in the case of Sanskrit, since pronunciation is strictly governed by the शिक्षा(/śikṣā/) (Pāṇini, 1938), a treatise on phonetics, schwa deletion is not observed.

2 List of works used in the speech corpus

- Mallinātha's commentary on KumāraSambhavam
- Mallinātha's commentary on Raghuvamśam
- Ādiśaṅkara's Bhaṣyam on Kaṭhopaniṣat
- Ādiśaṅkara's Bhaṣyam on Bhagavadgītā (Chapters 1-9)
- Ādiśaṅkara's Bhaṣyam on Brahmasūtram
- Yogasūtram Vyāsabhāṣya-sahitam
- Ṛṇvimuktiḥ by SaṃskṛtaBhāratī
- Āñjaneya-Rāmāyaṇam by SaṃskṛtaBhāratī
- Kathālaharī by SaṃskṛtaBhāratī
- Bālamodinī stories from SambhāṣaṇaSandeśa by SaṃskṛtaBhāratī
- Samarthaḥ Svāmī Rāmadāsaḥ by SaṃskṛtaBhāratī
- Yugāvatāraḥ by SaṃskṛtaBhāratī
- Prāstāvikam of Swāmī Aḍgaḍānanda's commentary on Bhagavadgītā
- ViśuddhaVedāntaSāraḥ by SaccidānendraSarasvatī
- Man-Kī-Bāt Sanskrit translation
- Lecture on Lilāvatī
- Extempore Discourse

2.1 Sources of Recorded Audios

- vedabhoomi.org
- https://archive.org/details/Anjaneya-rAmAyaNam
- https://archive.org/details/geethasb
- https://archive.org/details/bAlamodinI-01
- https://archive.org/details/kathA-laharI
- https://www.youtube.com/watch?v=LJGjfHHHBoQ
- https://sanskritdocuments.org/sites/manogatam/
- https://archive.org/details/YatharthGeetaSanskritAudio

2.2 Sources of Tools used for Recording, Cleaning and Transcribing the Audios

- ASR Voice Recorder https://play.google.com/store/apps/details?id=com.nll.asr
- Audacity https://www.audacityteam.org/
- oTranscribe https://otranscribe.com/

3 Computing Infrastructure

- GPU Model Name : GeForce GTX 1080 Ti
- GPU RAM : 12 GB
- CPU Model Name : CPU Intel(R) Xeon(R) Gold 5120 CPU
- Processor Speed : 2.20GHz
- System Memory : 256 GB
- CPU Cores : 56

References

Malhar Kulkarni, Chaitali Dangarikar, Irawati Kulkarni, Abhishek Nanda, and Pushpak Bhattacharyya. 2010. Introducing sanskrit wordnet. In *Proceedings on the 5th global wordnet conference (GWC 2010), Narosa, Mumbai*, pages 287–294.

Manomohan Ghosh Pāṇini. 1938. *Pāṇinīya Śikṣā or The Śikṣā Vedāṅga, Ascribed to Pāṇini*. University of Calcutta.

Richard G Salomon. 1996. Brahmi and kharoshthi. *The world's writing systems*, pages 373–383.

Richard Sproat. 2003. A formal computational analysis of indic scripts. In *International symposium on indic scripts: past and future, Tokyo*.