Same-Language Subtitles for low-resource Languages: A case of Bundelkhandi

Anirudh Pradhan*, Ayushi Pandey*, Divyansh Kushwaha, Akshita Tiwary, Vivek Seshadri
Karya
Bengaluru, Karnataka, India
{anirudh.pradhan, ayushi, divyansh, akshita, vivek}@karya.in

Abstract

Same-language subtitles enhance the viewing experience of audiovisual content for both the hearing impaired population and general audiences. However, while high-resource languages frequently benefit from automatic subtitling, subtitles are seldom available for content creators in regional languages, limiting audience engagement with independently produced media. To address this gap, this paper presents Project Saurakhi, a mobile-first platform dedicated to generating same-language subtitles in low-resource regional languages. Unlike standard desktop tools, Saurakhi is optimized for low-bandwidth environments, integrating automated speech recognition (ASR) with manual refinement by local human annotators. As a proof of concept, we curate a dataset comprising 63 hours of Bundelkhandi speech sourced from 207 YouTube videos across 19 content creators. Ultimately, Project Saurakhi aims to empower and train women participants in rural India to ethically generate gold-standard subtitles in their native languages, fostering digital inclusivity and enriching language technology research.

Keywords: corpus creation, low-resource languages, Bundelkhandi, Indian languages, conversational AI, speech recognition, YouTube data

1. Introduction

YouTube videos in Indian regional and vernacular languages attract active engagement from their native speakers. In India, affordable internet prices (Roy, 2019) allows for deep penetration of online media into semi-urban and rural areas. Therefore, the demand continues to increase for online content in regional languages from consumers of diverse socioeconomic groups (Mehta, 2020; Mohan and Punathambekar, 2019). In catering to this demand, YouTube videos, therefore, serve not only as forms of entertainment but also as important sites of under-resourced languages. However, these regional language videos are rarely accompanied by corresponding closed captions.

Closed captions, or same-language subtitles (SLS) were developed by the American School for the Deaf, when the production of silent films was taken over by "talkies" cinema (Boatner, 1981; Zárate, 2021). Since then, closed-captions remain an important basis of discussion in inclusive media (Mälzer and Wünsche, 2024; Yang et al., 2025; Gürkan, 2021). In addition to the hearing-impaired communities, closed-captions also benefit listeners with normal hearing. Recent surveys conducted on viewers of streaming over-the-top (OTT) platforms reveal that a significant majority (85% of Netflix users) (Ideal Insight, 2023), especially younger viewers, prefer to switch on captions while watching videos. The prevalence of this trend cannot simply be explained by improving intelligibility of dialogue, or comprehension in suboptimal listening conditions. Instead, closed captions are being employed to improve the overall dialogue understandability (Martinez et al., 2024), which reduces cognitive load and increases the quality of viewing experience. These studies underscore the importance of closed captions not only in accessibility but also in improving an overall viewership experience. The absence of subtitles, therefore, limits audience engagement with regional and minority-language content, which is often self-financed and independently produced. It also deprives the research communities in speech and language processing from accessing videos as a curated resource for minority languages.

To address this gap, we present Saurakhi, a video-annotation platform for regional language content on YouTube. The platform is integrated within the Karya app. The Karya app is a smartphone-based medium for ethical data collection that specialises in collecting data from low-income groups in rural India. The subtitles generated through Project Saurakhi aim to increase engagement and accessibility in Bundelkhandi content, provide paid employment to subtitle creators, and curate a high-quality, annotated video data that supports both accessibility and language technology research.

Section 2 outlines the need for creating such a platform, especially in the light of automatic subtitling options. Section 3 presents an overview of the data curation process and demonstrates the importance of YouTube content in the Bundelkhandi language. Section 4 presents a detailed description of the pipeline of the video annotation platform. Section 5 concludes the paper. Through this paper,

*equal contribution

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 3147–3153
11–16 May 2026.

we aim to describe content as an important regional language resource, and also describe the platform.

2. The need for a subtitling platform

2.1. Why can't we just ASR it?

Despite major advances in automatic speech recognition (ASR), human-generated subtitles still outperform automatic ones in terms of accuracy, readability, and user comprehension, even for English-language material. Empirical evaluations show that automatic captions produce significantly higher error rates than live human captions, often with three times as many errors per minute (Romero-Fresco and Fresno, 2023). In comprehension studies, viewers using professional human captions demonstrated better narrative understanding and engagement than those using automatic subtitles (Kim et al., 2023). Furthermore, in educational settings such as live lectures on Zoom or YouTube, human captions consistently achieved higher accuracy and semantic clarity than machine-generated alternatives (Kustritz et al., 2023). These findings confirm that while ASR-based subtitling systems are improving, human-generated captions remain the gold standard for accuracy and communicative effectiveness in English media.

2.2. Subtitling in low-resource languages

With the rise of the internet and accessibility, we see a global trend of community-generated content in regional languages. Consequently, there are initiatives in language preservation groups to facilitate these content creators. Specifically, the Projects like RISE UP (ris, 2023)(Revitalising Languages and Safeguarding Cultural Diversity), funded by HORIZON Europe, foster language revitalisation through audiovisual materials and subtitling. Similarly, grants are released for subtitling to increase the dissemination of audiovisual content in Catalonian (Llull, 2025).

Within the Indian context, accessibility targets of at least 50% of all audiovisual media have been laid out as a directive (Government of India, 2019). To advance this cause, the Billion Readers Initiative (BIRD) (, BIRD) has argued for 'born' accessibility, meaning that subtitles become germane to content before its release. Additionally, reading proficiency in Indian language scripts has been modelled and monitored through this initiative.

With this description, we aim to highlight that manual, gold-standard subtitling is an active field of both commercial and research interest. A smartphone-based medium, accessible with a simple internet connection, provides a feasible medium for providing annotations in regional languages. We next describe the Bundelkhandi language and show how Project Saurakhi can be used to advance its digital presence. While Bundelkhandi is chosen as a test case, the feasibility of the smartphone app-based medium will extend to other low-resource languages.

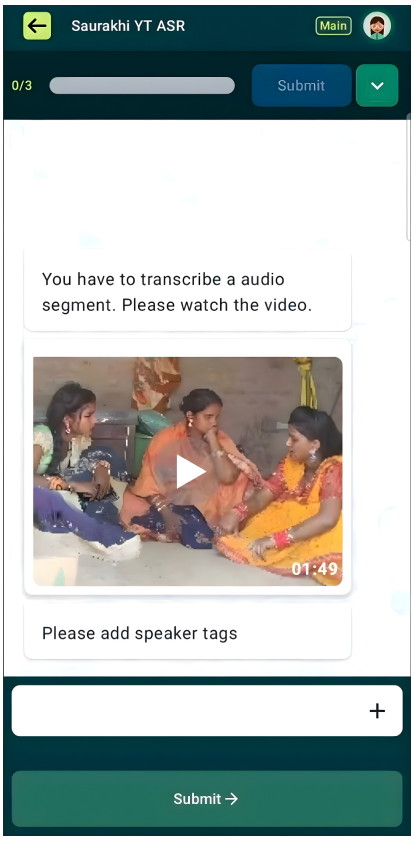

Figure 1: The user interface of the Saurakhi platform, showcasing the workflow for creating same-language subtitles for Bundelkhandi videos.

2.3. Bundelkhandi: a high-population, low-resource language

Project Saurakhi is aimed at those languages which have a large native speaker population, but very limited digital presence. Bundeli, or Bundelkhandi is an Indo-Aryan language, often regarded as a dialect of Western Hindi spoken predominantly in central India. Despite having a native speaker population of over 5.6 million, [1] the language does not enjoy an official status in India. Furthermore, its *digital* presence is significantly lower than other non-official languages of the Hindi belt (Bafna et al., 2022). For example, despite being unofficial languages, Bhojpuri, Maithili, and Awadhi have sizeable corpora (Kumar et al., 2023; Kumar and Choudhary, 2025) that are being employed by sev-

[1] Scheduled Languages in descending order of speaker's strength - 2011" (PDF). Registrar General and Census Commissioner of India. 29 June 2018.

Table 1: Distribution of content duration across YouTube creator channels. Data will be released after written consent from creators only.

Channel	Duration (HH:MM)	Avg. Length (MM:SS)	No. of Videos	Subscribers	%
Ashish Upadhyay	14:29	23:29	37	661k	22.8
Kakku Bhaiya	13:09	28:11	28	575k	20.7
Bihari Upadhyay	10:39	12:04	53	504k	16.8
Sunil Devraj	07:35	14:13	32	290k	12.0
Madarwas Boys	04:04	11:36	21	331k	6.4
Hariya Bhaiya	03:57	26:19	9	185k	6.2
Misspriya Bundeli	01:51	27:53	4	691k	2.9
13 Other Creators	07:16	19:58	23	12.65M	12.2
Total	**63:01**	**18:23**	**207**	**15.89M (approx)**	**100.0**

eral speech and natural language processing departments.

The only known digitalised resources of Bundelkhandi are the ASJP wordlist (Bafna et al., 2022) (Wichmann et al., 2018), obtained from the open-source poetry and folk resources. Additionally, a small corpus of 116 folk songs has been transcribed for genre classification in Bundelkhandi (Pandey and Dutta, 2014). A vernacular language newspaper, Khabar Lahariya, mentioned in (Pandey and Dutta, 2014) has discontinued its publication in vernacular languages, owing to reduced readership [2]. Finally, Bundelkhandi features in Project Vaani (Team, 2025), an initiative by Google to record phonological variation across Indian states.

While research communities in linguistics and computational spaces lag, content creators of Bundelkhand have turned to YouTube. Comedians, influencers and scriptwriters are actively releasing regional language content on YouTube, often maintaining a following of several thousand subscribers (see Table 1). These videos capture a range of vocabulary items, from vernacular to code-mixing patterns distinct from standard Hindi. As discussed before, these videos are not accompanied by subtitles.

This limits the engagement and accessibility of their wide consumer base. Moreover, the lack of transcribed data reduces the potential of ASR systems for automatic subtitling. Project Saurakhi aims to provide a platform for native speakers of Bundelkhandi to annotate this rich resource. The next section explains the video data in more detail, while Section 4 describes the details of the platform.

3. Corpus Design

3.1. Ethical concerns for YouTube data

Publicly available YouTube datasets have been a consistent source of language resources, especially in the analysis of “in the wild" speech and

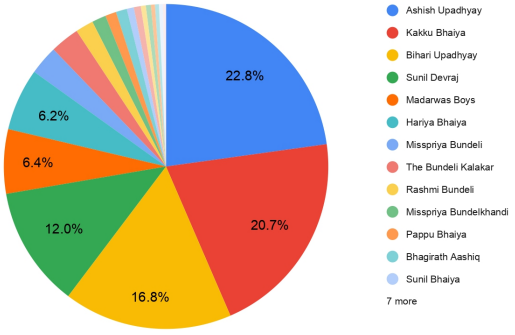

Figure 2: Distribution of content duration across YouTube creators in the Saurakhi Bundelkhandi Corpus.

language phenomena (Coats, 2023; Cooper, 2023) (Nagrani et al., 2017). A recent corpus on gender and sexual diversity (Siegert et al., 2025) identified that processing such data is justified in the public interest of fair, inclusive, and representative AI.

To establish a proof of concept for the Saurakhi platform, an audiovisual dataset in Bundelkhandi was collected using `yt-dlp` from selected YouTube channels. We acknowledge that while the fair dealing provisions of Section 52(1) of the Indian Copyright Act of 1957 permit limited use for research, this exception does not absolve the ethical requirement for creator consent. To respect the broader intellectual property of the creators, we extracted only a limited, non-substantial subset of videos from each channel, in adherence to the data minimization policies stated in (Kamocki et al., 2021).

It is crucial to clarify that the scope of this paper is strictly limited to demonstrating the architecture and functionality of the annotation platform. The crawled data is currently utilized exclusively in a closed environment for internal platform testing and validation. To address the ethical imperative of informed consent, our team will actively contact all

[2] information through email exchange

participating content creators to secure formal, written permission before the annotated Saurakhi Bundelkhandi Corpus is published, shared, or before any subtitles are uploaded back to their respective channels. This ensures our data practices remain entirely ethical and respectful of independent creators.

3.2. Three-Stage Technical Processing Pipeline

The technical workflow for preparing YouTube video data for annotation and transcription consists of three major stages: download and audio extraction, speaker diarization, and automatic speech recognition (ASR). Each stage contributes to converting raw audiovisual material into structured linguistic data suitable for further processing and analysis.

- **Stage 1: Download and Audio Extraction.** The source video is downloaded using the open-source tool `yt-dlp`. The audio stream is then extracted with `FFmpeg` and converted into a standardised 16-bit, 16 kHz mono-channel WAV format. This ensures compatibility with most ASR systems and provides a consistent audio quality baseline for subsequent processing.
- **Stage 2: Speaker Diarization.** To determine "who spoke when," an automatic speaker diarization model is applied to the continuous audio stream. The model segments the audio into discrete portions, each corresponding to a distinct speaker identity. This step is essential for handling multi-speaker conversations and ensuring accurate transcription alignment.
- **Stage 3: Automatic Speech Recognition (ASR).** Each speaker-labelled segment is transcribed using a pre-trained ASR model. In this study, we employ `wav2vec`-based architectures developed by AI4Bharat, trained on large-scale Indian language corpora. The resulting transcripts serve as the initial textual representation of the spoken content and are used for further annotation and quality refinement.

3.3. Current Data Mined

To date, we have built a scalable video pipeline and successfully processed **207 videos** for Bundelkhandi content creators. The videos were sourced from a total of **20 creators**, ensuring a diverse range of content and speaking styles. This effort has yielded a total of **63 hours** of audio data, of which **51 hours** is actual speech. The dataset is rich in speaker diversity, featuring an average of approximately 18 speakers per video. This collected data forms the initial version of the **Saurakhi Bundelkhandi Corpus**.

Figure 2 shows the distribution of content across different creators. The top two channels contribute 22.8% and 20.7% of the total duration, respectively, while the remaining content is distributed among 18 other creators, with the smallest 13 creators collectively contributing 12.2% of the corpus. Table 1 provides detailed statistics for each creator channel.

4. Subtitling Platform Description

To refine the preliminary machine-generated transcripts into a high-quality corpus, we developed a user-centric annotation platform integrated within the Karya app. This platform serves as the essential human-in-the-loop component bridging the gap between automated transcription and the creation of a gold-standard dataset. Unlike standard desktop-centric transcription tools that require expensive hardware and constant broadband connectivity, Saurakhi is uniquely optimized as a mobile-first application designed specifically for low-resource environments.

To accommodate the infrastructural realities of rural India, the platform features an offline-first architecture that allows annotators to download video batches when internet access is available and sync their completed transcriptions later. Additionally, the user interface is localized into regional languages to minimize the cognitive load on annotators.

The workflow begins where the technical pipeline (described in Section 3.2) concludes. Annotators are presented with video segments that have already been extracted from larger videos. As illustrated in Figure 1, the interface offers an intuitive yet powerful set of tools: a video player to review the segment, clear speaker labels (e.g., "Speaker 0," "Speaker 1") for tagging, and a dedicated text box for transcription.

Each annotator follows a systematic and sequential process as depicted in the flowchart in Figure 3. After watching the assigned video segment, the user performs two key actions in order:

1. **Speaker Tagging:** The annotator first identifies and labels the active speaker for the segment. This step ensures accurate attribution of speech before finalizing the transcription.
2. **Transcription and Subtitling:** With the speaker correctly tagged, the annotator transcribes the spoken content. This refines the preliminary ASR-generated text to ensure linguistic and contextual accuracy.

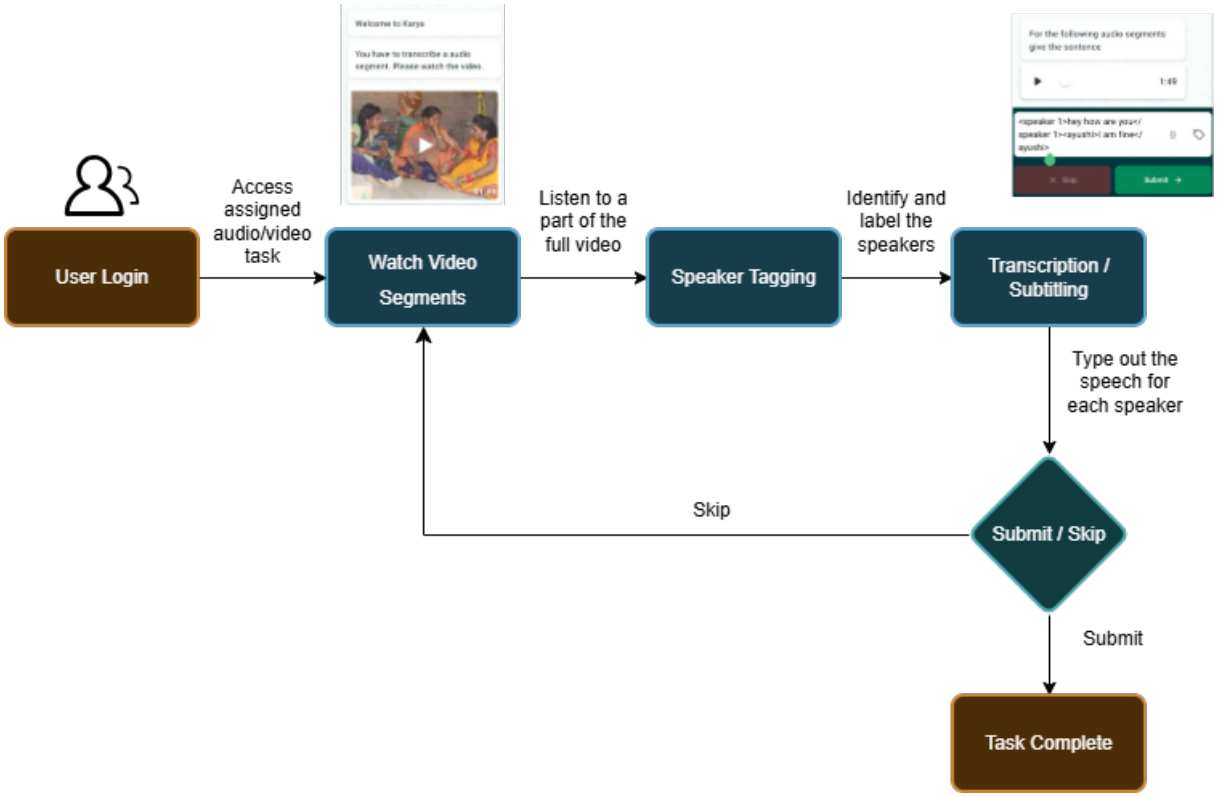

Figure 3: Flowchart describing the user annotation workflow.

After completing these steps, the annotator may either submit the finalized transcription or skip the segment if it is deemed untranscribable. To ensure the integrity of the data, we also integrated a robust Quality Assurance loop where a secondary layer of expert reviewers cross-verifies a subset of the submitted subtitles to establish a high Inter-Annotator Agreement. The subsequent subsection outlines the on-ground data collection roadmap in further detail.

5. Video annotation: collection and roadmap

Project Saurakhi is currently transitioning from content mining to the annotation phase, built upon Karya's core strength as an ethical data collector for low-income communities.

- Social Mission and Workforce Activation: First, we will identify a dedicated, paid cohort of 100 local women annotators across four key districts in the Madhya Pradesh Bundelkhand region: Damoh, Sagar, Jabalpur, and Garhakota.
- Guaranteed Paid Work and Scale: We are committing to the fair distribution of work. The 51 hours of Bundelkhandi speech will be allocated to ensure each annotator receives at least one hour of guaranteed paid annotation work.
- Gold-Standard Quality Assurance (QA): The large cohort of 100 annotators will be central to establishing the gold standard. We will use this size to establish robust Inter-Annotator Agreement (IAA) measures, ensuring the final same-language subtitles are accurate and reliable.

The resulting Saurakhi platform is available for all researchers, commercial service providers and also creative content developers who are interested in video annotation in regional languages. Once the annotation is complete, we will release the subtitles back onto YouTube, with written consent of the YouTube content creators. This is to support the content creators with increased engagement on their videos, many of whom are often independently produced.

6. Conclusion

In this paper, we described Project Saurakhi, which is a scalable and ethical platform for subtitle generation for audiovisual media. Centrally, we extend the functionality of a smartphone-based data collection medium to video annotation and source 63 hours of video data in the low-resource minority language Bundelkhandi on YouTube. This work provides a clear roadmap for addressing the critical lack of same-language subtitles, which currently limits accessibility and engagement for a signifi-

cant portion of India's diverse linguistic landscape. The platform's systematic manual subtitling process transforms ASR transcripts into gold-standard captions. Through these efforts, Project Saurakhi aims to create subtitled corpora for video content creators in low-resource languages of India and beyond.

7. References

2023. Revitalising languages and safeguarding cultural diversity (rise up) project fact sheet. CORDIS, European Commission. Grant agreement ID: 101095048. Accessed: 15 October 2025.

Niyati Bafna, Josef van Genabith, Cristina España-Bonet, and Zdeněk Žabokrtskỳ. 2022. Combining noisy semantic signals with orthographic cues: Cognate induction for the indic dialect continuum. In *Proceedings of the 26th Conference on Computational Natural Language Learning (CoNLL)*, pages 110–131.

Billion Readers (BIRD). 2025. A roadmap for accessible entertainment in india: Recommendations for access, literacy and language enhancement at population scale. Technical report, Billion Readers.

Edmund Burke Boatner. 1981. Captioned films for the deaf. *American annals of the deaf*, 156(5):520–525.

Steven Coats. 2023. Dialect corpora from youtube. *Language and linguistics in a complex world*, pages 79–102.

Christopher R Cooper. 2023. The identification of youtube videos that feature the linguistic features of english informal speech. *Applied Corpus Linguistics*, 3(3):100068.

Ministry of Information & Broadcasting Government of India. 2019. Office memorandum: Accessibility standards for persons with disabilities in television programmes. Technical Report No. 9/10/2012-BP&L (Vol-II), Ministry of Information & Broadcasting, Government of India. Issued under Sections 29(h) and 40 of the Rights of Persons with Disabilities Act, 2016; Accessed: 15 October 2025.

Ali Gürkan. 2021. Comparison of subtitling for the deaf and hard-of-hearing guidelines implemented accross countries. *Karamanoğlu Mehmetbey Üniversitesi Uluslararası Filoloji ve Çeviribilim Dergisi*, 3(2):249–278.

Ideal Insight. 2023. Netflix and captions?: A new study shows that 85% of netflix customers in the uk use subtitles while watching. Accessed: 21 November 2023.

Paweł Kamocki, Vladimir Popescu, Isabelle Gavanon, Camille Gaffiot, Khalid Choukri, and Valérie Mapelli. 2021. Elrc report on legal issues in web crawling (version 1.1). Revision date: 2021-02-09; originally published 2018; Accessed: 15 October 2025.

Hyunju Kim, Yan Tao, Chuanrui Liu, Yuzhuo Zhang, and Yuxin Li. 2023. Comparing the impact of professional and automatic closed captions on video-watching experience. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–6.

Ritesh Kumar, Meiraba Takhellambam, Bornini Lahiri, Amalesh Gope, Shyam Ratan, Neerav Mathur, and Siddharth Singh. 2023. Collecting speech data for endangered and under-resourced indian languages. In *Proc. 2nd Annual Meeting of the ELRA/ISCA SIG on Under-resourced Languages (SIGUL 2023)*, pages 14–18.

Shantanu Kumar and Narayan Choudhary. 2025. Maithili language technology: A survey. *Strength for Today and Bright Hope for Tomorrow Volume 25: 6 June 2025 ISSN 1930-2940*, page 160.

Margaret Root Kustritz, Ryan Rupprecht, and Perle Zhitnitskiy. 2023. Comparison of accuracy of machine-generated or human-generated captions of zoom live lectures in a comparative theriogenology course. *Clinical Theriogenology*, 15:52–56.

Institut Ramon Llull. 2025. Grants for the subtitling of audiovisual works produced and/or co-produced in catalonia (2025 call). Accessed: 15 October 2025.

Nathalie Mälzer and Maria Wünsche. 2024. Subtitling for people who are d/deaf or hard-of-hearing (sdh). In *Handbook of Accessible Communication*, pages 313–328. Frank & Timme, Berlin.

Helard Becerra Martinez, Alessandro Ragano, Diptasree Debnath, Asad Ullah, Crisron Rudolf Lucas, Martin Walsh, and Andrew Hines. 2024. Dialogue understandability: Why are we streaming movies with subtitles? *ArXiv*, abs/2403.15336.

Smith Mehta. 2020. Localization, diversification and heterogeneity: Understanding the linguistic and cultural logics of indian new media. *International Journal of Cultural Studies*, 23(1):102–120.

Sriram Mohan and Aswin Punathambekar. 2019. Localizing youtube: Language, cultural regions, and digital platforms. *International journal of cultural studies*, 22(3):317–333.

Ayushi Pandey and Indranil Dutta. 2014. Bundeli folk-song genre classification with knn and svm. In *Proceedings of the 11th International Conference on Natural Language Processing (ICON 2014)*, pages 133–138, Goa, India. NLP Association of India. Accessed: 11 October 2025.

Pablo Romero-Fresco and Nazaret Fresno. 2023. The accuracy of automatic and human live captions in english. *Linguistics and the Human Sciences (LANS–TTS)*, 22(1).

P Roy. 2019. Why india has world's cheapest mobile data. *BBC News*.

Hui Yang, Weiju Shen, Xinyan Ma, Bo Liu, and Yi Wang. 2025. Research on availability and reception of intralingual subtitles by the deaf and hard of hearing in china. *SAGE Open*, 15(3):21582440251365754.

Soledad Zárate. 2021. *Captioning and subtitling for d/deaf and hard of hearing audiences*. UCL Press.

Nagrani, Arsha and Chung, Joon Son and Zisserman, Andrew. 2017. *Voxceleb: a large-scale speaker identification dataset*. arXiv.

Siegert, Ingo and Marquenie, Jan and Grawunder, Sven. 2025. *Queer Waves: A German Speech Dataset Capturing Gender and Sexual Diversity from Podcasts and YouTube*. ISCA.

VAANI Team. 2025. *VAANI: Capturing the Language Landscape for an Inclusive Digital India (Phase 1)*. Indian Institute of Science (IISc) and ARTPARK.

Wichmann, S. and Brown, C. H. and Holman, E. W. and Dryer, M. S. and Ran, Q. 2018. *ASJP Database*. Cross-Linguistic Linked Data (CLLD). Accessed: 11 October 2025.