



BHASHINI and IISc Bengaluru : Text-to-Speech Model for Sourashtra Language

New AI-based Speech Technology Supports Digital Access, Preservation and Wider Use of a Low-Resource Indian Language

Posted On: 16 OCT 2026 2:32PM by PIB Delhi

In a significant step towards strengthening language technology for low-resource Indian languages, the Digital India BHASHINI Division (DIBD), under Digital India Corporation (DIC), Ministry of Electronics and Information Technology (MeitY), in collaboration with the Indian Institute of Science (IISc), Bengaluru, has developed Text-to-Speech (TTS) models for Sourashtra, a minority Indo-Aryan language spoken predominantly by the Sourashtra community in Tamil Nadu.

The initiative aims to expand the availability of speech technology for languages that have limited digital resources and support their use in digital applications, language learning, accessibility solutions, documentation and research.

The newly developed models use the VITS (Variational Inference with adversarial learning for end-to-end Text-to-Speech) architecture and are capable of generating Sourashtra speech directly from text. The models include male and female voices, with each voice available for text written in both Tamil script and Sourashtra script. This enables users to work with Sourashtra in the scripts used for the language.

The models have been trained using a dedicated Sourashtra speech corpus developed at the SPIRE Lab, Indian Institute of Science, Bengaluru. The corpus contains more than 28 hours of gender-balanced speech recordings, including recordings from one male and one female voice artist. The corpus includes composed sentences as well as sentences drawn from literature, with transcriptions available in both Sourashtra and Tamil scripts.

AIKOSH: BHASHINI's AIKOSH platform provides access to language datasets and resources that support the development of Indian language technologies. The SPICOR TTS 3.0 Sourashtra Corpus is now available on AIKOSH, offering 28+ hours of gender-balanced Sourashtra speech data developed at IISc Bengaluru. The corpus includes diverse and literary content in both Sourashtra and Tamil scripts, supporting the development of Sourashtra language technologies.

The male voice model was trained on approximately 10,000 utterances, while the female voice model was trained on approximately 11,400 utterances. The speech corpus and associated text data were carefully prepared and cleaned to support the development of

the TTS models.

The models generate audio at 22,050 Hz and operate directly on character-level text, eliminating the need for a separate phonemizer. The four models comprise two speaker voices, each available in Tamil-script and Sourashtra-script versions.

Sourashtra, spoken by around 2.48 lakh people according to the 2011 Census, has a relatively limited digital presence and fewer available speech technology resources. The development of these models goes beyond language preservation by helping bring Sourashtra into the digital ecosystem. The technology can benefit the Sourashtra-speaking community by enabling access to digital content in their language, support students and language learners through speech-enabled learning tools, and improve accessibility for users who rely on spoken digital content. At the same time, the availability of the models and underlying speech resources provides researchers, developers and institutions with a foundation to build further Sourashtra language technologies and applications across education, accessibility and other public-facing digital services.

The models can support applications including reading Sourashtra text aloud, language learning, accessibility solutions and research in low-resource speech synthesis, while contributing to the broader development of digital resources for Sourashtra and other underrepresented Indian languages.

Through collaborations with academic and research institutions, BHASHINI continues to strengthen the ecosystem for Indian-language datasets, speech technologies and AI-based language solutions, particularly for languages that remain underrepresented in the digital domain.

About Digital India BHASHINI Division (DIBD)

The Digital India BHASHINI Division (DIBD), under the Digital India Corporation (DIC), Ministry of Electronics and Information Technology (MeitY), is building a national ecosystem for AI-powered multilingual digital inclusion. BHASHINI supports 36 Indian text languages, 23 Indian voice languages and 35 international languages, and enables the integration of Language AI capabilities across government and public-facing digital services. Its ecosystem includes technologies for translation, speech recognition, speech synthesis, transliteration and other language services, helping make digital information and services more accessible across India's linguistic diversity.

Vinod Kumar / Kanishk Sharma

(Release ID: 2324289) Visitor Counter : 457

Read this release in: Telugu , Urdu , हिन्दी , Tamil , Kannada

Share

