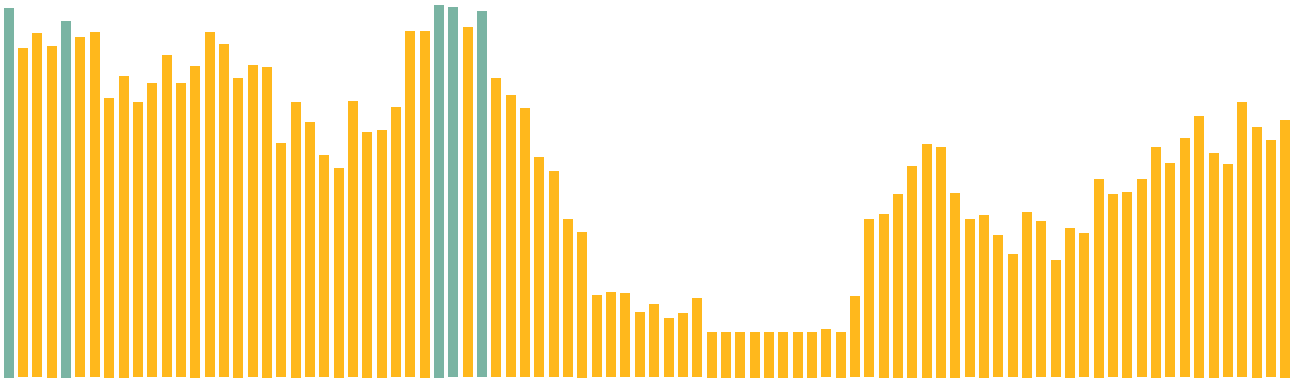


A voice envelope: 5 hours of Soliga speech, recorded on the edge.

Frugal Voice

Offline community governed speech models

Building a federated, frugal speech model with the Soliga tribe of Karnataka, India.

A language preserved on the edge —
low cost, low connectivity, community owned.

FRUGAL AI HUB at UNIVERSITY OF CAMBRIDGE Judge Business School

in collaboration with

IIT Dharwad
ज्ञानेन विकासः

Frugal Voice: Offline community governed speech models for the Soliga tribe in Karnataka

Authors

Arjuna Sathiaseelan **(lead author / corresponding)**
Technology Lead, Frugal AI Hub · Visiting Fellow, Cambridge Judge Business School
a.sathiaseelan@jbs.cam.ac.uk

Nandha Sathiaseelan
Researcher, Frugal Voice Project

Sarbani Banerjee Belur
Co-Founder & Community Research Lead, Saving Voices Project · IIIT Dharwad

Prashant Bannulmath
Collaborator, IIIT Dharwad

Sunil Saumya
Collaborator, IIIT Dharwad

Shruti Maralappanavar
Collaborator, IIIT Dharwad

Deepak K T
Collaborator & Assistant Professor, IIIT Dharwad

Serish Venkata Gandikota
Co-founder, Frugal AI Hub · Visiting Fellow, Cambridge Judge Business School
s.gandikota@jbs.cam.ac.uk

Elizabeth Osta
Co-founder, Frugal AI Hub · Visiting Fellow, Cambridge Judge Business School
e.osta@jbs.cam.ac.uk

Jaideep Prabhu
Professor of Marketing & Nehru Professor of Indian Business, Cambridge Judge Business School
j.prabhu@jbs.cam.ac.uk

About the Frugal AI Hub

The Frugal AI Hub at Cambridge Judge Business School is a research and practice hub advancing the principles, methods and adoption of Frugal AI — intelligent systems that are high performing yet resource efficient, and that can be understood and governed by the institutions and communities that deploy them. The Hub works with industry, public sector and civil society partners across sectors including fintech, healthcare, education, sustainability and community-built AI for linguistic and cultural preservation.

Partner institution

This work was carried out in collaboration with the **Indian Institute of Information Technology, Dharwad (IIIT Dharwad)**, an Institute of National Importance established in 2015 under the Ministry of Education, Government of India. IIIT Dharwad provided speech and signal processing expertise, fieldwork support in Karnataka, and on-the-ground partnership with the Soliga community throughout this project.

Abstract

Low resource and tribal languages in India are at acute risk of digital extinction because contemporary language technologies overwhelmingly target high resource languages and centralized, cloud hosted large models that are expensive to train, deploy, and adapt. These constraints are particularly visible for the Soliga community, a Dravidian language speaking tribe in the Biligiri Rangaswamy Hills of Karnataka, whose language has no widely used script and for which digital resources remain scarce.

In this paper we present a socio–technical framework and prototype for building low cost voice models with and for the Soliga tribe using offline, edge based federated learning. The framework combines community governed data collection and consent processes, on-device training of compact ASR and keyword spotting models, and an intermittent connectivity aggregation protocol that runs on low cost edge hardware.

Our experiments show that, even with only 5 hours of Soliga speech, a centralized Wav2Vec 2.0 baseline achieves a word error rate (WER) of 37.95% and character error rate (CER) of 11.11%, and that a federated frugal architecture trained on the same small corpus can approach this performance (WER 44.20%, CER 15.80%) while keeping data local and under community control, and delivering practically useful keyword spotting (F1 0.84) for health-related phrases.

> *“Frugal AI is not only about doing more with less. For tribal communities like the Soliga, it is about doing it locally, in their own language, on their own terms.”*

1. Introduction

Low resource and tribal languages in India are at acute risk of digital extinction because contemporary language technologies overwhelmingly target high resource languages and centralized, cloud hosted large models that are expensive to train, deploy, and adapt. Centralized automatic speech recognition (ASR) and text-to-speech (TTS) pipelines require large amounts of labelled data, stable connectivity, and significant compute resources, all of which are misaligned with the infrastructural realities and socio-economic conditions of many tribal communities in India. These constraints are particularly visible for the Soliga community, a Dravidian language speaking tribe concentrated in the Biligiri Rangaswamy Hills of Karnataka, whose language has no widely used script and for which digital resources remain scarce.

Several initiatives have explored AI tools to support tribal languages in India, including translation and speech technologies for languages such as Santali, Mundari, and Gondi, typically in partnership with tribal research institutes and native speakers. While impactful, most such systems rely on centralized infrastructure and cloud-based deployment, which complicates long term sustainability and local ownership in bandwidth constrained and intermittently connected environments. In parallel, work on frugal AI argues that marginalized communities need context-specific solutions that are resource-aware by design and that minimize data, compute, and energy while maximizing local control and socio-economic value. Yet there remains little empirical evidence on whether such frugal architectures can deliver practically useful speech performance for purely oral, low resource languages under strict connectivity and governance constraints.

Federated learning (FL) offers a promising technical and governance mechanism to reconcile collaborative model training with strong data sovereignty and privacy requirements. Instead of uploading raw audio to a central server, FL enables edge devices to train local models and share only model updates, which can then be aggregated into a global model without moving sensitive speech data off device. This paradigm is particularly attractive for tribal communities, where distrust of external actors, historical extractivism in data collection, and infrastructural constraints make conventional data-hungry AI pipelines inappropriate and unsustainable.

1.1 Research questions

In this paper we ask two questions. **First**, can a community governed federated architecture for Soliga approach the performance of a centralized ASR baseline while operating entirely on low cost, intermittently connected edge devices? **Second**, how do Soliga community members perceive such models in terms of data sovereignty, instrumental value for accessing services, and their role in cultural and linguistic preservation? Addressing these questions requires not only technical evaluation but also a socio-technical design that gives the community meaningful control over data and models.

1.2 Contributions

Our contributions are threefold:

Methodology	A socio-technical methodology for co-designing federated learning pipelines with a tribal community that emphasises data governance, consent, and linguistic documentation, including a community data council with veto power over model sharing.
Architecture	A frugal edge architecture for offline training and intermittent aggregation of speech models tailored to Soliga, deployable on Raspberry Pi-class hardware, with explicit design for weeks-long offline operation.
Empirical evaluation	An evaluation showing that a centralized Wav2Vec 2.0 ASR baseline trained on 5 hours of Soliga speech achieves a WER of 37.95% and a CER of 11.11%, and that a federated frugal architecture with 10 simulated edge clients can approach this performance (WER 44.20%, CER 15.80%) while preserving data locality and delivering practically useful keyword spotting performance (F1 0.84) for health-related phrases.

Throughout, we deliberately restrict ourselves to a corpus of only 5 hours of Soliga speech. This reflects realistic documentation budgets for non-orthographic tribal languages and allows us to evaluate whether a federated, frugal architecture remains viable under such extreme data scarcity.

2. Related work

2.1 Speech technologies for tribal and low-resource Indian languages

There is a growing body of work on documenting and developing speech technologies for under-represented Indian tribal languages, including Soliga, Lambani, Mundari, and Kui. Recent efforts have collected several hours of tribal speech data and used modern deep learning architectures to build ASR and TTS systems, often transliterating into dominant regional scripts such as Kannada for Soliga and Lambani or Odia for Mundari and Kui because these languages lack standardized native scripts. These projects demonstrate the feasibility of high quality speech synthesis and recognition for languages with small speaker populations, but they typically rely on centralized training workflows and do not explicitly address community data governance or edge deployment.

National scale initiatives such as Bhashini and related platforms aim to preserve tribal languages and make government services accessible via AI-powered translation and speech tools. These systems integrate text-to-speech and translation models for multiple tribal languages and are deployed via cloud infrastructure to reach users across devices. While impactful for accessibility, this paradigm tends to centralize both the data and the control over model evolution, raising questions around local ownership, adaptability to offline settings, and long term sustainability for communities with limited connectivity.

Standard ASR systems struggle with high degrees of phonetic variation and dialectal diversity within a single tribal language. They often force communities to adopt external scripts that do not fully capture the language's sound system, and they frequently neglect community participation, which leads to mistrust and poor data quality. These gaps motivate approaches that treat language as lived heritage rather than as a commodity and that embed governance and cultural objectives directly into model design.

2.2 Federated learning for speech and low-resource settings

Federated learning has been widely studied as a privacy-preserving mechanism for training models across distributed clients without centralizing raw data. In speech applications, prior work has applied FL to keyword spotting and personalized ASR and has shown that federated optimisation can match or approach centralized performance when communication and aggregation are carefully designed. Most of this literature assumes relatively stable connectivity and focuses on application level personalization for well-resourced languages rather than on first generation models for low resource languages in offline community settings.

2.3 Frugal AI and community-built models

Frugal AI has been proposed as a broader design philosophy that goes beyond efficiency and advocates for AI systems that are co-designed with resource constraints and social goals at the forefront. Rather than retrofitting large, general-purpose models to new domains, frugal AI emphasizes small, task-specific models, minimal data use, and context-sensitive architectures that can be owned and maintained locally. Community-built language technologies for tribal groups increasingly involve native speakers directly in data collection, annotation, and evaluation, often through on-site recording campaigns and participatory design workshops. Our work builds on these approaches and intersects them with federated learning so that data locality is preserved and Soliga speakers can iteratively improve the models without ceding control over their voice data.

3. The Soliga community: geographic and socio-technical context

3.1 Geography and settlement patterns

The Soliga (also spelled Sholaga) are an indigenous, forest dwelling community (Adivasi) residing primarily in and around the Biligiri Rangaswamy Temple (BRT) Tiger Reserve in the Chamarajanagar district of southern Karnataka, India. Geographically, the BR Hills hold immense ecological significance, serving as a vital bridge between the Eastern and Western Ghats. The Soliga people have historically coexisted with this rich biodiversity.

Their settlements are organized into small, decentralized hamlets known as *podus*, which are dispersed deep within the forest canopy or along its fringes. Because of strict forest conservation regulations and the challenging terrain, these *podus* are geographically isolated. They frequently lack paved road access and are located at substantial distances from primary markets, schools, and healthcare facilities.



Figure 1. *The Soliga community broadcasting station, situated within a forest podu in the BR Hills.*

3.2 Demography and language

Demographically, the Soliga community comprises several tens of thousands of individuals across the region. They speak the Soliga language, a morphologically rich Dravidian language that shares linguistic roots with Kannada and Tamil. Crucially for the design of speech technologies, Soliga is an exclusively oral language with no widely adopted, standardized native script. When textual representation is required, the language is typically transliterated into the dominant regional Kannada script. The purely oral nature of the language, combined with a small speaker population and the absence of vast textual corpora, places Soliga at acute risk of digital extinction in an era dominated by text heavy, data hungry foundational AI models.

3.3 Livelihoods and autonomy

Historically, the Soliga were nomadic hunter gatherers who practiced shifting cultivation. Over the past century, state conservation policies and the declaration of wildlife sanctuaries restricted these traditional practices. Today, their local economy is heavily dependent on small scale, rain fed agriculture and the sustainable harvesting of non timber forest products (NTFPs) such as wild honey, lichen, gooseberry (amla), and medicinal plants.

The community possesses a strong, deeply ingrained culture of collective governance and advocacy. Notably, the Soliga have a long history of successfully asserting their indigenous rights. Following the passage of the landmark Forest Rights Act (FRA) of 2006, they became the first tribal community in India to secure formal community forest rights within the core area of a tiger reserve. This legal and cultural legacy of fighting for physical resource autonomy directly informs their contemporary demands for digital and data sovereignty.

3.4 Infrastructural and socio-technical realities

The geographic isolation of the *podus* translates into severe infrastructural constraints that dictate the design boundaries of any technological intervention. The region suffers from intermittent and highly fluctuating electricity, with many households relying on improvised power sharing arrangements or off grid solar alternatives. Cellular network coverage is weak, fragmented, or entirely non existent in many hamlets, rendering traditional cloud centric, always connected mobile applications virtually unusable.

Despite these infrastructural barriers, there is a growing, instrumental aspiration for digital connectivity among younger, bilingual Soliga adults. Consequently, there is a distinct need for localized, Soliga language voice interfaces. Yet, given the infrastructural realities and the community's right to self determination, such systems cannot rely on centralized architectures. They must be offline first, highly compact, and capable of operating on low power edge hardware.

4. Methodology

Our methodology combines community-centred fieldwork with a frugal, federated training pipeline that operates on low cost edge devices in offline settings.

4.1 Community engagement and governance

We began with a co-design process involving Soliga community leaders, local NGOs, and linguists familiar with Soliga. The aims were to define acceptable use cases for voice models, such as access to health information and documentation of oral narratives, to agree on data governance rules, and to identify trusted local institutions to host aggregation nodes. Informed consent protocols were developed in the Soliga language, and participants could opt into or out of specific use cases.



Figure 2. *The research team with Soliga community leaders during a co-design and consent workshop.*

Data ownership was assigned to the community through a local cooperative structure that formally controls any sharing of model artefacts beyond the community. We adopted a three-tier participation framework:

Tier 1 — Speakers	Volunteer Soliga speakers who contribute voice recordings on community-owned devices, with opt-in/opt-out at the use-case level.
Tier 2 — Facilitators	Local facilitators who manage devices, explain consent, conduct quality checks and monitor weekly training cycles.
Tier 3 — Data Council	Community data council that oversees when and how global models are updated, deployed, or shared externally — with explicit veto power.

4.2 Data collection protocol

Speech data consists of read and semi-spontaneous utterances in Soliga, recorded on low cost Android smartphones, Raspberry Pis and low cost laptops with USB microphones. We curated a balanced prompt set that includes everyday phrases related to health, education, and public services; short narratives and folk stories recorded with elder speakers; and phonetically rich sentences adapted from existing Soliga linguistic documentation and Kannada corpora mapped to Soliga phonology.



Figure 3. *Voice recording sessions with Soliga speakers using low cost smartphones, laptops and USB microphones.*

All recordings are stored locally on devices in compressed form (16 kHz mono, Opus/FLAC) and tagged with minimal metadata such as age bracket, gender, and consent tier. No personally identifying information is attached to the audio. Basic on-device quality checks, such as signal-to-noise ratio thresholds and duration limits, filter out unusable samples before they enter any training cycle.

4.3 Model tasks and labels

Given the relatively small data budget and the need for simple and interpretable functionality, we focus on two primary tasks. **First**, we train a small-vocabulary keyword spotting (KWS) model for phrases related to health and service access — for example equivalents of “doctor”, “hospital”, and “medicine” in Soliga. **Second**, we train a character-level or subword-level ASR model that maps Soliga speech to a Kannada-script transliteration suitable for linguists and bilingual community workers. To reduce annotation burden, we adopt a semi-supervised regime in which unlabeled audio segments receive pseudo-labels from the current global model and are later verified or corrected by annotators.

4.4 Federated training protocol

We adopt a cross-device federated learning setup tailored to intermittent connectivity and low-power hardware. Client devices include smartphones and edge boxes held by community members and local centres; the server is a community controlled mini-server located at a local NGO or school and reachable via local Wi-Fi or periodic physical synchronisation.

Each training round proceeds as follows. The edge server maintains the current global model parameters, which are compressed and versioned for distribution. When devices come into Wi-Fi range, for example during weekly community meetings, they receive the latest model snapshot via local file transfer. Devices then train the model locally on stored audio with on-device feature extraction, data augmentation, and mini-batch updates for a fixed number of steps under energy and time constraints. They compute model deltas, apply 8-bit quantization and gradient sparsification to reduce communication size, and upload updates during the next connectivity window. The server performs a weighted FedAvg update over participating devices and uses simple norm and loss-based checks to detect malformed or outlier updates.

4.5 Ethical and privacy considerations

The federated design keeps raw audio recordings on devices controlled by community members or trusted local institutions, which reduces the risk of external misuse. Safeguards include on-device encryption of stored audio with device-level PINs managed by local facilitators, explicit opt-out mechanisms that trigger local deletion of recordings and exclusion from future training rounds, and model-card-style documentation co-written with the community that specifies intended use, limitations, and prohibited uses such as surveillance or profiling. The community data council reviews new deployment proposals, including any external research collaborations, and can veto sharing of models or derived artefacts.

5. System architecture

5.1 High-level architecture

The overall architecture comprises three main layers:

Device layer	Low cost Android phones and Raspberry Pi-class boxes used for data collection, local training, and on-device inference.
Community edge layer	A small, always-on aggregation server hosted at a local institution that stores global model versions and aggregates updates.
External support layer	Optional university or NGO servers that may receive anonymized model snapshots (not raw updates) for analysis — only with explicit community approval.

5.2 Model architecture choices

For keyword spotting, we use a compact convolutional neural network operating on log-Mel spectrogram patches with depthwise-separable convolutions to reduce parameter count and computation. For ASR, we adopt a small encoder-decoder architecture: the encoder consists of 3-4 layers of bidirectional GRUs or a small Conformer-style stack, and the decoder is a character-level or subword-level attention decoder that predicts Kannada script tokens representing Soliga speech.

Both models are further compressed via post-training 8-bit quantization for on-device inference, optional pruning of low-magnitude weights after initial training phases, and knowledge distillation from a larger teacher model trained in a controlled environment where a small amount of de-identified surrogate data can be shared under community approved conditions.

5.3 Centralized ASR baseline: Wav2Vec 2.0

To establish a strong centralized upper bound, we finetune a Wav2Vec 2.0 model on the full Soliga speech corpus. Wav2Vec 2.0 is a self-supervised speech representation model comprising a convolutional feature extractor, a Transformer-based encoder, and a quantization module. During pre-training, the model learns from large amounts of unlabeled speech via a masked contrastive prediction task. In the fine-tuning phase, the pre-trained encoder is adapted to ASR by adding a task-specific projection head and optimizing the entire model using a connectionist temporal classification (CTC) loss.

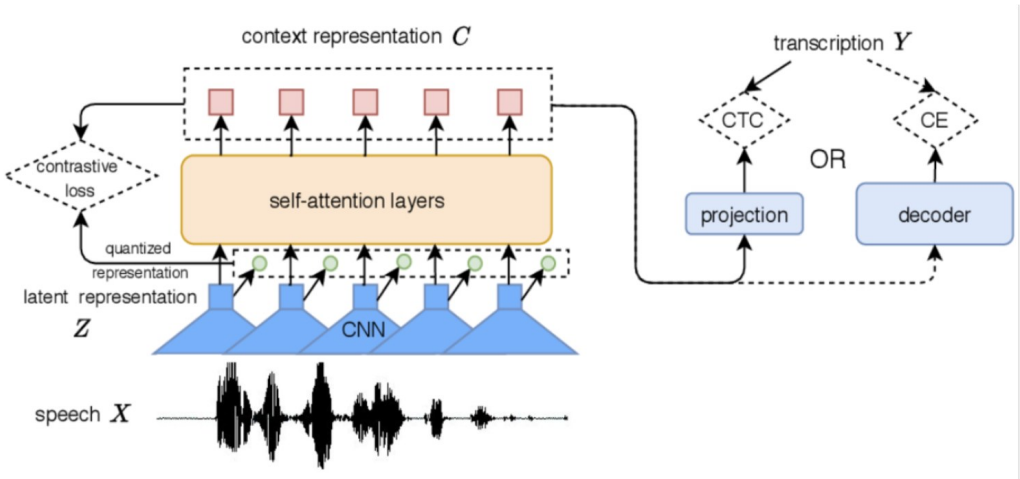


Figure 4. *Schematic of the Wav2Vec 2.0 architecture used as the centralized ASR baseline.*

5.4 Offline operation and delayed synchronisation

A key architectural requirement is resilience to prolonged offline periods. Devices are expected to operate without any network connectivity for weeks, collecting and locally training on new audio data. The synchronisation module maintains a local queue of un-uploaded updates, each tagged with the originating global model version. Upon reconnection, it uploads pending updates, allowing the aggregator to apply them in order or discard obsolete deltas when version skew becomes too large. Devices are encouraged, via the interface, to synchronise at community gathering points such as weekly markets or school events where the edge server is accessible.

6. Experimental setup and results

6.1 Dataset and partitions

The speech corpus was collected in collaboration with the Soliga community under the governance and consent protocols described in Section 4. Two complementary datasets were assembled: an ASR corpus of read and semi-spontaneous utterances on low cost devices, and a TTS corpus recorded in more controlled conditions to support future synthesis experiments.

Corpus	Purpose	Utterances	Duration
ASR corpus	Automatic speech recognition	10,000	5 hours
TTS corpus	Text-to-speech synthesis	4,200	4 hours

Table 1. *Soliga speech corpus composition.*

Split	Unique words	Avg. sentence length (words)
Train	13,455	4.55
Validation	2,692	4.55
Test	2,741	4.59

Table 2. *Dataset statistics: unique words and average sentence length by split.*

The relatively large vocabulary for a 5-hour corpus reflects the morphological richness of Soliga as a Dravidian language and the phonetically diverse prompt design. The corpus was partitioned at the speaker level using a 70/15/15 train/validation/test split to avoid speaker overlap.

6.2 Baselines

We compare three experimental conditions: (i) a **centralized Wav2Vec 2.0 baseline** finetuned on all 5 hours of Soliga audio on a university server with CTC loss; (ii) a **local-only** baseline in which each simulated edge client trains compact models on its own ~0.5-1 hour partition with no knowledge transfer; and (iii) the **federated frugal model (ours)** with 10 simulated edge clients and intermittent FedAvg aggregation at the community edge server, initialised from a small Wav2Vec 2.0-distilled seed and exchanging 8-bit quantized, sparsified deltas.

6.3 Evaluation metrics

We evaluate keyword spotting using F1-score and accuracy on a held-out test set. For ASR, we report character error rate (CER) and word error rate (WER) on held-out Soliga utterances in Kannada transliteration. We also track resource metrics: on-device training time per round on Raspberry Pi 4-class hardware, estimated energy usage, and per-update communication volume after quantization and sparsification.

6.4 Quantitative results

Model	KWS F1	CER (%)	WER (%)	Train time (min)	Update size (KB)
Centralized (Wav2Vec 2.0)	0.91	11.11	37.95	N/A	N/A
Local-only baseline	0.62	29.40	68.50	8	N/A
Federated frugal (ours)	0.84	15.80	44.20	12	340

Table 3. *Performance and resource metrics across experimental conditions.*

The centralized Wav2Vec 2.0 baseline achieves a WER of 37.95% and a CER of 11.11% on 1 hour of held out Soliga test data, establishing an upper bound for performance when all data is centralized. The federated frugal model achieves a WER of 44.20% and a CER of 15.80%, compared to 68.50% and 29.40% for the local-only baseline. The gap of approximately 6.3 WER points between the federated model and the centralized baseline suggests that the proposed architecture can approach centralized performance while preserving data locality. The CER gap is even narrower, indicating that many federated errors arise at the level of word boundaries or morphology rather than at the character level.

The federated KWS model attains an F1 of 0.84, close to the centralized 0.91 and substantially higher than the local-only 0.62. The average per-round local training time of 12 minutes on Raspberry Pi 4-class hardware is compatible with overnight or background training, and the average compressed update size of 340 KB compares favourably with federated keyword spotting deployments that report on the order of several megabytes per client over 100 rounds.

6.5 Positioning within the Soliga and tribal ASR landscape

Language	Model	Train data	WER (%)	CER (%)
Soliga	W2V2 (centralized)	5 h	37.95	11.11
Soliga	Whisper Large-V2 (fine-tune)	~11 h	22.09	—
Soliga	Federated frugal (ours)	5 h (dist.)	44.20	15.80
Lambani	Whisper Large-V2 (fine-tune)	~11 h	49.61	—
Kannada	IndicWav2Vec	large-scale	26.94	—
Garhwali	W2V2-BERT (fine-tune)	low-resource	49.30	19.30

Table 4. *ASR performance comparison for Soliga and related tribal languages (figures are indicative; corpora and evaluation setups differ across rows).*

Notably, the federated frugal Soliga model outperforms the Whisper Large-V2 result for Lambani despite using less data and a decentralized training protocol, and its CER of 15.80% is competitive with Wav2Vec-BERT finetuned on Garhwali. These comparisons suggest that the federated frugal approach does not catastrophically degrade performance relative to centralized alternatives while providing the advantages of data sovereignty and offline deployment.

7. Qualitative and socio-technical findings

We conducted qualitative interviews and focus groups with 24 Soliga participants from three villages, including eight elder speakers, ten younger bilingual adults, and six community leaders and NGO staff. Sessions were conducted in Soliga with translation support from Tier 2 facilitators. Thematic analysis revealed three main themes.

Data sovereignty and trust	Instrumental value	Cultural preservation
Participants consistently valued that voice recordings did not leave community controlled devices and that the data council retained veto power over any external sharing. Elders contrasted the present project with earlier extractive recording campaigns.	Younger bilingual speakers saw KWS as a tool for interacting with government health services, NGOs, and educational resources — for example navigating automated health lines or conveying medication names to Kannada-speaking clinicians.	Elders emphasised preserving folk stories, medicinal knowledge, and ritual songs and favoured scenarios that recorded narratives rather than only transactional phrases. ASR-based Kannada transliterations were seen as a possible bridge for intergenerational transmission.

Importantly, the KWS F1 of 0.84 was perceived as practically sufficient for constrained interactions, particularly when bilingual facilitators could intervene in case of recognition errors. Elders also expressed concerns about misrepresentation of Soliga phonology in Kannada script and argued for community controlled annotation and validation workflows.

8. Discussion and future work

The empirical results indicate that the federated frugal architecture represents a viable middle ground between a fully centralized ideal and a purely local baseline. The federated frugal model incurs an absolute WER increase of approximately 6.3 points relative to the centralized Wav2Vec 2.0 baseline, a roughly 16% relative degradation, but reduces WER by 24.3 absolute points compared to the local-only condition. In the Soliga context, where any functioning ASR model is a first, we view this as an acceptable tradeoff for complete data locality and community control.

The centralized Wav2Vec 2.0 baseline and the federated frugal model both perform substantially worse than the best reported Whisper Large-V2 result for Soliga (22.09% WER). Much of this gap is attributable to model capacity and pretraining: Whisper Large-V2 has on the order of 1.5 billion parameters and is pretrained on hundreds of thousands of hours of multilingual audio, whereas our student models are two to three orders of magnitude smaller. Importantly, Whisper finetuning requires datacentre-class GPUs and continuous connectivity to centralized infrastructure, which are unavailable in the Biligiri Rangaswamy Hills. The frugal architecture deliberately accepts a 16-22 WER-point gap in exchange for deployment on sub-$50 Raspberry Pi class hardware that can operate entirely offline and under community governance.

8.1 Limitations

Our study has several limitations. First, the ASR dataset comprises only 5 hours of speech collected from a modest number of speakers. While typical for first-round documentation efforts on non-orthographic tribal languages, it is small by conventional ASR standards. Second, the federated experiments simulate 10 clients and intermittent connectivity on a single server; future deployments should validate the protocol on actual devices in the field with realistic power cycles and participation patterns. Third, the qualitative component covers 24 participants from three villages and is therefore exploratory rather than representative of all Soliga sub-groups.

8.2 Future work

Future work proceeds in three directions. First, we plan to extend the framework to additional tribal languages with contrasting phonological and governance contexts, such as Lambani and Gondi. Second, we aim to integrate simple on-device interfaces for community-driven evaluation and feedback — for example allowing speakers to flag mis-recognised tokens or culturally inappropriate outputs that can be fed back into training. Third, we are interested in hybrid training regimes in which a small, community approved subset of anonymised data is used to train larger teacher models in a controlled environment, while students remain small and fully edge deployable. This could narrow the gap to Whisper-scale performance without compromising data sovereignty.

Beyond technical extensions, there is a need for sustained engagement with policymakers and standards bodies to formalise community rights over voice data and models, drawing on indigenous data governance frameworks such as OCAP and recent debates on AI and Indigenous data sovereignty.

9. Conclusion

In this paper we have documented our experiences building Frugal Voice, a socio-technical framework and prototype for community governed, offline speech models for the Soliga tribe in Karnataka. By combining federated learning, frugal model design, and a community data council, the system allows Soliga speakers to build and control ASR and keyword spotting models on low cost hardware without relinquishing their voice data. Our results show that a federated frugal architecture can approach the performance of a centralized Wav2Vec 2.0 baseline and remain competitive with other low resource Indic ASR systems while operating entirely at the edge. Qualitative findings highlight that the governance structure, not only the technical architecture, is crucial for building trust and aligning speech technologies with both instrumental and cultural goals. We hope this work contributes to a broader agenda of frugal, community controlled AI for endangered and marginalized languages.

References

Baevski, A., Zhou, H., Mohamed, A., & Auli, M. (2020). wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. *Advances in Neural Information Processing Systems*, 33.

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2023). Robust Speech Recognition via Large-Scale Weak Supervision. *Proceedings of the 40th International Conference on Machine Learning*.

Rajkhowa, T., Spandana, S., et al. (2024). Evaluating the Efficacy of Large Acoustic Models for Documenting Non-Orthographic Tribal Languages in India. *Proceedings of LREC-COLING 2024*.

Spandana, S., Rajkhowa, T., et al. (2023). Documenting the Soliga Language: Linguistic and Community Perspectives. *International Journal of Dravidian Linguistics*, 1-20.

Dhuriya, S., Gupta, Y., et al. (2022). IndicWav2Vec: A Multilingual Speech Model for Low-Resource Indic Languages. *Proceedings of Interspeech*.

Srivastava, A., et al. (2024). Frugal AI: Design Principles for Resource-Constrained AI Systems. *Frugal AI Initiative White Paper*.

Leroy, D., Coucke, A., Lavril, T., Gisselbrecht, T., & Dureau, J. (2019). Federated Learning for Keyword Spotting. *ICASSP 2019*.

Kairouz, P., McMahan, H. B., Avent, B., et al. (2021). Advances and Open Problems in Federated Learning. *Foundations and Trends in Machine Learning*, 14(1-2), 1-210.

McMahan, H. B., Moore, E., Ramage, D., Hampson, S., & Aguera y Arcas, B. (2017). Communication-Efficient Learning of Deep Networks from Decentralized Data. *Proceedings of AISTATS*.

Ministry of Electronics and Information Technology. (2023). National Mission on Natural Language Translation (BHASHINI). *Government of India*.

Kumar, R., Singh, P., et al. (2025). Dialect Matters: Cross-Lingual ASR Transfer for Low-Resource Indic Varieties. *arXiv preprint 2601.04373*.

Rajkhowa, T., Chowdhury, A. R., Bannulmath, P., Deepak, K. T., & Prasanna, S. R. M. (2024). Optimizing Direct Speech-to-Text Translation for Un-orthographic Low-Resource Tribal Languages Using Source Transliterations. *Proceedings of the 25th International Workshop on Spoken Language Technologies for Under-Resourced Languages (COCOSDA)*.

Vaswani, A., Shazeer, N., Parmar, N., et al. (2017). Attention Is All You Need. *Advances in Neural Information Processing Systems*, 30.

Frugal AI Hub
at Cambridge Judge Business School

Trumpington Street
Cambridge CB2 1AG
United Kingdom

https://frugalai.org

Contact

Arjuna Sathiaseelan
Lead author and corresponding contact
a.sathiaseelan@jbs.cam.ac.uk

FRUGAL AI HUB at UNIVERSITY OF CAMBRIDGE Judge Business School

in collaboration with

IIIT Dharwad
ज्ञानेन विकासः