POSITION PAPER

Rethinking Multilingual Speech for Under-Represented Languages:

Tokenisation, Training, and Evaluation Beyond the English-Centric Paradigm

Dr. Arjuna Sathiaseelan
Founder and Principal Investigator, Saving Voices
March 2026

Abstract

Indigenous and other under-represented language communities face an acute and often irreversible crisis: as the last fluent speakers of thousands of languages age, the window for documentation and preservation closes permanently. Current multilingual speech models systematically fail these communities, not merely because of data scarcity but because the foundational design choices of dominant architectures reflect English-centric assumptions about sound, script, and structure. This paper argues that meaningful progress requires rethinking tokenisation, multilingual training, and evaluation from first principles. It advocates byte- and morpheme-aware tokenisation regimes suitable for oral-tradition languages with no standardised orthography, self-supervised acoustic pre-training on typologically diverse corpora paired with frugal fine-tuning strategies, and linguistically grounded evaluation protocols that go beyond word error rate to capture tonal, morphological, and socio-linguistic nuance. Drawing on the Saving Voices Project's work with the Soliga community of Karnataka, India, an end-to-end research workflow and a set of design principles are presented to guide partners who build, deploy, and govern voice models in indigenous and other under-represented communities.

1. Introduction

Of the approximately 7,000 languages spoken on earth today, roughly 40 per cent are endangered, many surviving only in the memory of a small number of elders (Hammarstrom et al., 2023). When a language disappears, it carries with it an entire worldview: ecological knowledge accumulated over centuries, oral histories, spiritual traditions, and ways of understanding human experience that exist in no other form. The loss is irreversible. Voice technology offers one of the most powerful tools available for documentation and revitalisation, enabling communities to record, archive, and teach their languages at scale. Yet fewer than 100 of the world's 7,000 languages are meaningfully represented in commercial automatic speech recognition (ASR) systems, and indigenous languages fare worst of all.

The Saving Voices Project was founded on the conviction that this failure is not inevitable. It is the consequence of design choices made for English and a handful of well-resourced languages,

applied without adaptation to radically different linguistic contexts. Indigenous languages present challenges that compound every stage of the conventional speech pipeline: they are frequently tonal, agglutinative, or polysynthetic; they may have no standardised orthography or writing tradition; their speaker populations may be measured in hundreds rather than millions; and the urgency of documentation is extreme, as speaker communities are often elderly and shrinking.

This paper takes a clear position: the English-centric speech pipeline is not a neutral baseline from which indigenous and other under-represented languages can simply be fine-tuned. It is a structurally biased system whose assumptions about subword units, acoustic frames, orthographic regularity, and evaluation metrics actively impede progress. Incremental adaptation is insufficient; principled redesign is required. The paper identifies three axes along which redesign must occur: (1) tokenisation and subword representation, (2) multilingual acoustic pre-training and adaptation, and (3) evaluation frameworks. For each axis the paper describes the problem, surveys recent evidence, and articulates concrete design principles grounded in real-world experience with indigenous language communities.

Scope. The term under-represented language denotes any language for which: (a) fewer than 10 hours of transcribed speech are publicly available; or (b) standard benchmarks yield word error rates above 30 per cent for state-of-the-art models; or (c) the language exhibits structural properties including tonality, complex morphology, and non-Latin script that cause systematic failure modes in English-derived architectures. This paper gives particular attention to indigenous languages, which satisfy all three conditions and face the additional challenge of imminent speaker loss. The category also includes non-indigenous languages with millions of speakers (e.g., Hausa, Amharic, Burmese) that nonetheless remain technologically marginalised (Conneau et al., 2021).

2. The Distinctive Challenges of Indigenous Languages

While all under-represented languages face barriers in current speech technology, indigenous languages present a distinct cluster of challenges that require targeted attention. These challenges are not merely technical; they are socio-linguistic, ethical, and logistical in nature, and they interact in ways that make generic low-resource solutions inadequate.

2.1 Oral Traditions and the Absence of Orthography

Many indigenous languages have no established writing system, or have multiple competing orthographic conventions developed by different missionary, governmental, or community actors. This creates a fundamental problem for ASR systems, which require a consistent mapping between sound and symbol. In the absence of a standardised orthography, the research team must make consequential decisions: which script to adopt, how to represent sounds that exist in the language but not in any available script, and how to handle the fact that different community members may write the same word in different ways. These decisions are not merely

technical; they have political and cultural implications for the community, and communities must be centrally involved in making them (Himmelmann, 2006).

For purely oral languages, the documentation pipeline must begin before ASR is even attempted. Fieldwork linguists must work with community members to develop a consistent phonological transcription system, train local annotators, and create the first written records of a language that may never previously have been written. This is laborious, specialist work that cannot be automated, and it must precede any machine learning effort.

2.2 Extreme Data Scarcity and Speaker Demographics

Many endangered indigenous languages have fewer than a few hundred fluent speakers, and those speakers are often elderly. This creates a data scarcity problem of a qualitatively different order from that faced by, say, a low-resource national language with millions of speakers but limited digital presence. The total available speech may be measured in minutes rather than hours, and the window for collection is closing. Standard approaches to data augmentation, which assume that the training distribution can be diversified by recording more speakers, may be inapplicable when the entire speaker community is a handful of individuals.

Speaker demographics also affect acoustic diversity in ways that matter for model robustness. If all available recordings feature elderly speakers, the model will perform poorly for younger speakers attempting to learn and use the language, precisely the population whose engagement is most critical for revitalisation (UNESCO, 2003).

2.3 Polysynthetic and Complex Morphological Structures

Many indigenous languages of the Americas, Australia, and parts of Africa and Asia are polysynthetic: a single word may express what English requires an entire sentence to convey, incorporating information about subject, object, tense, aspect, evidentiality, and spatial orientation into a single phonological unit. This presents extreme challenges for tokenisation strategies derived from English, where words are relatively short and morphologically simple. A polysynthetic word submitted to a standard BPE tokeniser may be split into ten or more fragments, none of which correspond to meaningful linguistic units, destroying all morphological signal.

2.4 Tonality and Non-Standard Phoneme Inventories

Many indigenous languages, particularly in sub-Saharan Africa, Southeast Asia, and parts of the Americas, are tonal: the pitch contour of a syllable determines its meaning. Others have phoneme inventories that include sounds absent from any European language: clicks (as in the Khoisan languages and Nguni languages of southern Africa), retroflex consonants, complex vowel harmony systems, or distinctive features such as phonation type (breathy, creaky, or modal voice) that standard acoustic features do not capture well. Models pre-trained primarily on European languages lack the representational capacity to distinguish these sounds reliably (Gut, 2008).

2.5 Code-Switching and Language Contact

Indigenous communities rarely exist in linguistic isolation. Speakers typically also speak a dominant regional or national language (Hindi, Spanish, Portuguese, French, Indonesian, and so on), and natural speech frequently involves switching between the indigenous language and the dominant language within a single utterance. Current ASR systems handle monolingual speech poorly for indigenous languages; they handle code-switched speech almost not at all. Any system intended for real-world use in an indigenous community must be designed to handle code-switching gracefully, rather than treating it as an error.

2.6 The Urgency of the Documentation Window

Unlike low-resource national languages, where the speaker population is stable and the documentation challenge is primarily one of resource allocation, many indigenous languages face an imminent and irreversible documentation deadline. When the last fluent speaker of a language passes away without being recorded, that language is lost forever. This creates an ethical imperative that does not exist for other categories of under-represented language: imperfect documentation now is categorically better than perfect documentation planned for later. Research and deployment timelines must reflect this urgency.

Case Study: The Soliga Tribe, Karnataka, India: A Saving Voices Pilot

The Soliga are among the oldest indigenous forest-dwelling communities of Karnataka, India, inhabiting the Biligiri Rangaswamy Temple Wildlife Sanctuary for generations. Their language, Soliga (also written Sholiga), belongs to the Dravidian family and is spoken by approximately 50,000 to 60,000 people, though the number of fully fluent speakers, particularly among younger generations, is declining rapidly.

The Saving Voices Project conducted its inaugural voice documentation study with the Soliga community, developing a frugal AI pipeline for indigenous language preservation. The project faced all five challenges described in this section simultaneously: Soliga has no standardised writing system, a small and demographically skewed speaker population, complex Dravidian morphology, a phonological system that diverges significantly from Indo-European norms, and pervasive code-switching with Kannada and Tamil.

Community champions were trained as voice data collectors, ensuring that data collection was conducted by and for the Soliga community rather than imposed from outside. Natural speech, oral traditions, and daily language use were recorded under a community-consent framework. Lightweight ASR and text-to-speech models were built specifically to run on low-powered devices, ensuring that technology remained accessible regardless of infrastructure constraints.

The Soliga methodology is the direct foundation for the design principles articulated in this paper. It is designed to be replicable across any indigenous language community globally, providing a scalable blueprint that prioritises community ownership, linguistic rigour, and frugal, deployable technology.

3. The Structural Misalignment of Current Speech Pipelines

The dominant speech pipeline consists of three stages: (1) acoustic feature extraction, typically log-Mel filterbanks; (2) a sequence model mapping acoustic frames to subword tokens; and (3) a language model or decoder that converts token sequences to text. Each stage embeds English-centric assumptions that compound to produce poor performance on structurally different languages and catastrophic failure on indigenous languages (Prabhavalkar et al., 2023).

3.1 Acoustic Feature Extraction

Standard log-Mel filterbanks were designed to approximate the human auditory system as characterised through English and European-language data. The frequency resolution and temporal window sizes optimised for English consonant-vowel transitions are poorly suited to the fine-grained pitch discrimination required for tonal languages. In Yoruba, for example, the same phoneme sequence /ba/ realised with high, mid, or low pitch corresponds to entirely different words (Gut, 2008). Standard filterbanks discard precisely the high-frequency temporal modulations that carry this information. Similarly, the click consonants of Nguni languages (isiZulu, isiXhosa) and the pharyngeal fricatives of Semitic languages occupy acoustic regions undersampled by standard filterbank configurations. For indigenous languages with phonation-type distinctions or complex tonal systems, this failure is often total: the acoustic representation simply does not encode the features that carry meaning.

3.2 Subword Tokenisation

Byte-pair encoding (BPE), the dominant subword tokenisation strategy, learns merge rules from a training corpus. When that corpus is English-dominated, the resulting vocabulary is deeply biased (Sennrich et al., 2016). Agglutinative languages, where a single word may encode what English expresses in an entire phrase, are shattered into fragments that lack phonological or morphological coherence. For polysynthetic indigenous languages, the problem is still more severe: a single word form may be split into a dozen arbitrary substrings that correspond to nothing in the linguistic structure of the language. For languages with no established orthography, the tokenisation layer must handle ad hoc transcription conventions that differ between annotators, further degrading consistency.

3.3 Language Model Priors

Decoder language models impose priors on which character sequences are plausible. When these models are pre-trained on English text, as is universal in commercial systems, they systematically penalise valid surface forms of under-represented languages. The problem is not merely data quantity: a language model that has never encountered a Dravidian script, a Mayan syllabary, or an Ethiopic syllabary will assign near-zero probability to any indigenous language hypothesis, regardless of acoustic evidence. Beam search then systematically selects incorrect but familiar alternatives (Radford et al., 2023).

3.4 Evaluation Metrics

Word error rate (WER), the standard benchmark, is poorly calibrated for morphologically rich languages and almost meaningless for polysynthetic indigenous languages. In a language where a single word encodes an entire proposition, a single word substitution may represent total communicative failure, yet WER scores it identically to an innocuous single-word error. For tonal languages, WER ignores tone marking entirely in most implementations, rendering high apparent accuracy compatible with unintelligible output (Schuller et al., 2013). These metric failures mean that the field has no reliable way to measure whether it is making progress for indigenous language speakers.

4. Tokenisation: Byte- and Morpheme-Aware Regimes

Tokenisation is the first point of contact between raw text and a neural model. For indigenous and other under-represented languages, poor tokenisation is not merely inefficient; it actively destroys the linguistic signal that would allow models to generalise. A two-tier approach is advocated: byte-level fallback for robustness, combined with language-aware morpheme-level segmentation where resources permit.

4.1 The Case for Byte-Level Tokenisation

Byte-level models, which treat the UTF-8 encoding of text as their atomic unit, are inherently script-agnostic. Any language that can be encoded in Unicode is representable without out-of-vocabulary tokens (Everson and Lillie, 2001). For indigenous languages where the orthography is still being developed or where multiple conventions coexist, this robustness is especially valuable: byte-level models will not catastrophically fail when they encounter an unfamiliar diacritic or a community-specific character. Recent work demonstrates that byte-level models can match or exceed subword models on morphologically complex languages while eliminating tokenisation-induced fragmentation (Xue et al., 2022).

4.2 Morpheme-Aware Segmentation

Where morphological analysers exist, even simple rule-based ones, morpheme-aware segmentation substantially outperforms BPE. A morpheme is a linguistically meaningful unit; splitting on morpheme boundaries preserves the signal that acoustic and language models can exploit. For agglutinative and polysynthetic languages, morpheme-aware tokenisation can reduce vocabulary size by an order of magnitude while increasing coverage, because morphemes are the productive units of word formation (Kamper et al., 2022). In the Soliga project, working with community linguists to develop a morpheme inventory before training began was one of the highest-return investments made in the entire pipeline.

A practical pipeline is proposed: (1) attempt morpheme segmentation using available tools; (2) fall back to character-level segmentation for unknown forms; (3) use byte-level encoding as the

final fallback for unseen scripts or mixed-script text. This cascading strategy ensures that no input is unrepresentable while preferring linguistically motivated units when they are available.

4.3 Design Principles for Low-Resource Tokenisation

DP-T1	Adopt byte-level tokenisation as the default fallback for any language lacking a curated subword vocabulary. This is especially important for indigenous languages where orthographic conventions are still being established.
DP-T2	Invest in lightweight morphological analysers before investing in additional data collection. A rule-based morphological analyser built by a community linguist in two weeks can yield larger performance gains than 100 additional hours of transcribed speech.
DP-T3	Evaluate tokenisation quality independently from end-to-end ASR quality. Metrics such as fertility (average tokens per word), out-of-vocabulary rate, and morpheme boundary recall should be tracked and reported alongside WER.
DP-T4	Involve the community in orthographic decisions. Where no standard orthography exists, the choice of script and spelling conventions must be made with and by the community, not imposed by technologists. These choices shape all downstream modelling.

5. Multilingual Acoustic Pre-Training and Frugal Fine-Tuning

The dominant paradigm for low-resource ASR is to fine-tune a large pre-trained model on a small target-language dataset. For indigenous languages, this approach fails doubly: the pre-trained model's representations were shaped by languages typologically distant from the target, and the fine-tuning dataset may be so small (minutes, not hours) that standard fine-tuning leads to immediate overfitting. A different strategy is required, one designed for the extreme low-resource, high-typological-distance regime.

5.1 Rethinking the Pre-Training Distribution

Whisper, the current dominant open-weight ASR model, was trained on 680,000 hours of weakly supervised speech data scraped from the internet (Radford et al., 2023). The language distribution of this data heavily reflects online content creation patterns, which are themselves skewed towards high-income, English-speaking populations. Indigenous languages are almost entirely absent. A more principled approach is to construct a pre-training corpus that is intentionally typologically diverse, sampling languages to cover the major structural families (tonal, agglutinative, polysynthetic, isolating, fusional) and phonological inventories including click consonants, tone systems, and non-European vowel and consonant inventories.

The FLEURS benchmark and the Massively Multilingual Speech project represent steps in this direction, demonstrating that typologically balanced pre-training yields substantially better zero-shot and few-shot transfer to low-resource languages than English-dominated baselines (Fleurs et al., 2022). Initiatives such as ALFFA (African Languages in the Field: Frugal ASR) and

the Endangered Languages Project provide additional resources and methodological guidance for indigenous language documentation.

5.2 Self-Supervised Pre-Training Objectives

Supervised pre-training requires transcribed data, which is the very resource that is scarce for indigenous languages. Self-supervised objectives, which learn from unlabelled audio, are therefore particularly valuable (Baevski et al., 2020). Three objectives that have proven effective in low-resource settings are highlighted:

- **Masked Acoustic Modelling (MAM):** Randomly mask segments of the acoustic input and train the model to predict the masked frames. This forces the model to learn contextual acoustic representations without requiring labels, and is effective even on very small datasets.
- **Contrastive Predictive Coding (CPC):** Train the model to distinguish true future frames from distractors. CPC representations have been shown to capture phonological structure in a language-agnostic manner, making them a valuable initialisation for indigenous language models.
- **Cross-lingual Acoustic Alignment:** Train on language pairs where one language is well-resourced and one is not, using typological similarity to transfer representations. This is particularly effective when a related language (such as a dominant regional language spoken by the same community) has more resources.

For the Soliga project, cross-lingual transfer from Kannada and Tamil, both Dravidian languages with substantially more resources, provided a substantially better initialisation than transfer from English-based models, confirming that typological relatedness matters more than raw data volume.

5.3 Frugal Fine-Tuning Strategies

Given the extreme scarcity of labelled data for indigenous languages, fine-tuning strategies must be designed to extract maximum value from very small datasets. Three approaches are advocated.

5.3.1 Parameter-Efficient Fine-Tuning (PEFT)

Methods such as LoRA (Low-Rank Adaptation) and adapter layers allow a small number of parameters to be fine-tuned while the pre-trained weights remain frozen (Hu et al., 2022). For languages with fewer than 10 hours of data, and especially for indigenous languages where data may be measured in minutes, full fine-tuning consistently leads to catastrophic forgetting. PEFT methods preserve cross-lingual representations while adapting to the target language's phonological and prosodic characteristics. In the Soliga project, adapter-based fine-tuning on fewer than 3 hours of recorded speech produced models that community members judged usable for documentation purposes.

5.3.2 Linguistically Informed Data Augmentation

Augmentation strategies must be chosen with linguistic awareness. For tonal indigenous languages, pitch-preserving augmentation is essential: augmentation that shifts pitch contours risks generating acoustically plausible but linguistically incorrect training examples. Speaker normalisation augmentation is particularly valuable when initial data are collected from a small number of elderly speakers, as it diversifies the acoustic distribution to better cover younger or learner speakers.

5.3.3 Semi-Supervised Learning and Elder-Speaker Prioritisation

Given that unlabelled recordings (e.g., existing archival recordings, community radio broadcasts, or ceremony recordings) are often more abundant than carefully transcribed speech, semi-supervised approaches can multiply the effective training data. For indigenous languages where elder speakers are the most fluent but also the most demographically at risk, active learning should explicitly prioritise coverage of elder-speaker acoustic characteristics, even at the expense of other forms of diversity.

5.4 Design Principles for Pre-Training and Adaptation

DP-A1	Prioritise typological relatedness over data volume when selecting pre-training corpora. For Dravidian languages, a Dravidian pre-training corpus is more valuable than an English one, even if the English corpus is 100 times larger.
DP-A2	Use parameter-efficient fine-tuning by default for all indigenous language datasets. Full fine-tuning is almost never appropriate at the data scales available for endangered languages.
DP-A3	Treat elder speakers as the highest-priority data source. Their speech is the most linguistically authentic, the most irreplaceable, and the most urgent to record. Augmentation strategies should be designed to extend, not replace, elder-speaker data.
DP-A4	Design models to run on low-powered, community-owned devices. A model that requires cloud infrastructure is not accessible to remote indigenous communities. Frugality is not a compromise; it is a design requirement.

6. Linguistically Grounded Evaluation

Evaluation is not merely a measurement problem; it is a values problem. The choice of metrics determines what properties of a model are deemed important, which failures are treated as acceptable, and which communities receive attention when resources are allocated. For indigenous language communities, the stakes of poor evaluation are especially high: a model that appears adequate on standard benchmarks but fails on elder-speaker speech or culturally significant registers may cause active harm by creating false confidence (Ardila et al., 2020).

6.1 The Inadequacy of Word Error Rate

WER measures the edit distance between a hypothesis transcript and a reference transcript, normalised by reference length. For indigenous languages, WER is a poor proxy for several distinct reasons:

- Polysynthesis: In a polysynthetic language where a single word encodes an entire proposition, a single word substitution may represent total communicative failure. WER treats this identically to a trivial function-word substitution.
- Tone blindness: Most WER implementations treat tone diacritics as optional or normalise them away. A model that transcribes a tonal word with the wrong tone scores a perfect WER despite producing a semantically wrong or unintelligible transcription (Gut, 2008).
- Orthographic inconsistency: For languages with no standardised orthography, the reference transcription itself may vary between annotators. WER computed against such references is unreliable, potentially penalising correct transcriptions that use a different but equally valid spelling.
- Cultural register insensitivity: Indigenous languages often have ceremonial, narrative, or elder registers that differ substantially from everyday speech and that carry the greatest cultural value. WER calculated on everyday speech reveals nothing about performance in these registers.

6.2 A Multi-Dimensional Evaluation Framework

A five-dimensional evaluation framework is proposed, designed to capture the linguistic properties relevant to indigenous and other under-represented languages. These dimensions are not mutually exclusive and should be reported jointly.

Dimension	Challenge	Metric / Approach
Phonological Fidelity	Does the transcription preserve phonologically contrastive features, including tone, phonation type, and non-European phonemes?	Tone Error Rate (TER); Distinctive Feature Error Rate (DFER); phoneme-level confusion matrices for non-European phoneme inventories.
Morphological Integrity	Does the transcription preserve morpheme boundaries and morphological content, including polysynthetic structures?	Morpheme Error Rate (MER); morpheme boundary F1; semantic role preservation across morphological transformations.
Semantic Adequacy	Is the meaning of the utterance preserved in the transcription?	Semantic textual similarity between hypothesis and reference; downstream task performance on transcribed text.

Socio-Linguistic Appropriateness	Does the model perform equitably across elder and younger speakers, ceremonial and everyday registers, and code-switched speech?	Disaggregated WER/MER by speaker age, register, and code-switching frequency.
Practical Usability	Can community members use the system to accomplish real documentation and revitalisation tasks?	Task completion rate in community-led user studies; perceived usefulness assessed by community members in their own language.

6.3 Tonal Evaluation in Detail

Tone is a phonemic feature in approximately 70 per cent of the world's languages by speaker count and is particularly prevalent among indigenous languages of Africa, Southeast Asia, and the Americas. Tone Error Rate (TER) is proposed as a standard metric, defined as the proportion of toned syllables for which the transcribed tone category differs from the reference, after alignment. For languages with lexical tone (e.g., Yoruba, Mandarin, Igbo, and many indigenous languages), TER should be reported alongside WER as a primary metric. For languages with grammatical tone, TER should be supplemented by a grammatical function preservation score (Ludeling and Kyto, 2008).

6.4 Community-Centred Evaluation

Technical metrics are necessary but insufficient. The ultimate test of a speech model for an indigenous community is whether community members find it useful for the purposes they care about: teaching the language to children, preserving the speech of elders, accessing cultural materials, and conducting daily life in their language. Structured community evaluation protocols are advocated, distinct from user experience studies conducted by model developers, in which community members assess model outputs according to criteria they themselves define.

Community evaluation must explicitly surface failure modes that technical metrics cannot capture: whether the model handles ceremonial or narrative registers, whether it correctly transcribes elder speech, and whether it manages the code-switching patterns that characterise natural speech in indigenous communities. Models that systematically fail on elder speech, who are the primary carriers of linguistic knowledge, should not be deployed without remediation, regardless of aggregate benchmark performance (Muller et al., 2021).

6.5 Design Principles for Evaluation

DP-E1	Report Tone Error Rate as a primary metric for any tonal language. Never report WER alone; doing so obscures the most linguistically significant failure mode.
DP-E2	Disaggregate all metrics by speaker age, register, and code-switching frequency before any deployment decision. Models that perform well on aggregate but fail for elder speakers are not suitable for indigenous language documentation.

DP-E3	Conduct at least one round of community-led evaluation before deployment. This evaluation should be conducted in the indigenous language, compensated at a locally appropriate rate, and produce findings that are owned by the community.
DP-E4	Define deployment thresholds in terms of practical documentation utility, not benchmark scores. A model that an elder speaker finds unusable should not be deployed, regardless of its WER.

7. An End-to-End Research Workflow

The design principles articulated in the preceding sections are most effective when implemented as part of a coherent, sequential workflow. This section describes the workflow developed through the Saving Voices Project's experience with the Soliga community, adapted for general application to indigenous and other under-represented language communities.

Phase 1: Language and Community Audit (Weeks 1 to 4)

Before any modelling work begins, a thorough audit of the target language's linguistic properties and the community's needs and constraints must be conducted. This audit should document the typological profile (tone system, morphological type, script status, orthographic conventions if any, and degree of standardisation), the existing resource inventory (any transcribed speech, archival recordings, text materials, morphological knowledge, and related-language resources), the use-case requirements (what the community actually wants to do with the technology), and the community governance structure (who has authority to make decisions, what data sovereignty requirements apply, and what consent processes are culturally appropriate).

For indigenous languages with no orthography, Phase 1 must also include a community consultation on script and spelling conventions. No modelling work should begin without this consultation completed and its outputs documented.

Phase 2: Data Strategy and Collection (Weeks 4 to 16)

For indigenous languages, data collection is often the most resource-intensive and time-critical phase. Community champions should be trained as voice data collectors, following the Saving Voices model: local people who understand the cultural context, have the trust of elder speakers, and can conduct recording sessions in a culturally appropriate manner. The recording protocol should prioritise elder speakers, narrative and ceremonial registers, and naturally occurring code-switched speech. All data collection must operate under a community-consent framework, with data ownership vested in the community from the outset.

Phase 3: Tokenisation Design (Weeks 8 to 16)

Tokenisation strategies should be developed and evaluated in parallel with data collection. Byte-level tokenisation serves as the baseline. Where community linguists can provide morphological guidance, morpheme-aware segmentation should be developed and evaluated.

Tokenisation quality metrics, including fertility, OOV rate, and morpheme boundary recall, should be reported before proceeding to acoustic modelling.

Phase 4: Pre-Training and Adaptation (Weeks 12 to 32)

Select a pre-trained model with the highest typological overlap with the target language. For Dravidian languages, prefer models with Dravidian pre-training; for Bantu languages, prefer models with Bantu pre-training; for isolating tonal languages of Southeast Asia, prefer models that include tonal pre-training data. Apply PEFT fine-tuning with linguistically informed data augmentation. All fine-tuned models must run on community-owned, low-powered devices (Hu et al., 2022).

Phase 5: Multi-Dimensional Evaluation (Weeks 28 to 40)

Conduct systematic evaluation across all five dimensions of the framework described in Section 6. Disaggregate by speaker age, register, and code-switching frequency. Conduct an independent community evaluation round with compensated community evaluators. Define deployment go/no-go criteria based on practical documentation utility, agreed with the community in Phase 1.

Phase 6: Deployment and Ongoing Stewardship (Week 36 onwards)

Deploy with continuous monitoring of production performance. Establish a community feedback channel, staffed by community champions, and a regular review cadence. All models and training data must be version-controlled and remain under community ownership. Evaluation results, including failures and limitations, must be published openly and shared with the community in their language.

8. Governance and Indigenous Data Sovereignty

For indigenous language communities, governance is not a secondary concern to be addressed after the technical work is complete; it is the foundation on which all technical work must rest. The historical relationship between indigenous communities and outside researchers has too often been extractive: knowledge and cultural materials have been collected, published, and commercialised without meaningful community involvement or benefit. Voice technology must not replicate this pattern.

8.1 The CARE Principles for Indigenous Data Sovereignty

The CARE Principles for Indigenous Data Governance (Carroll et al., 2020) provide a framework specifically designed for the context of indigenous communities and research partnerships. All Saving Voices project work is conducted in accordance with these principles:

- **Collective Benefit:** Data ecosystems should be designed and function in ways that enable indigenous peoples to derive benefit from the data. Models trained on community

speech must return tangible value to that community, whether as free access to deployed tools, capacity building, or other forms agreed with the community.

- **Authority to Control:** Indigenous peoples' rights and interests in indigenous data must be recognised and their authority to control such data must be affirmed. This means data ownership agreements that vest primary ownership in the community, not the collecting organisation.
- **Responsibility:** Those working with indigenous data have a responsibility to share how those data are used and to support indigenous peoples' capacity to use the data for their own benefit. This includes translating model documentation into the indigenous language and training community members to maintain and update models.
- **Ethics:** Indigenous peoples' rights and wellbeing should be the primary concern at all stages of the data and research lifecycle. This includes the right to refuse data collection, to withdraw consent, and to require deletion of data at any time.

8.2 Data Sovereignty in Practice

Data sovereignty for indigenous language speech means: (a) data ownership agreements that vest primary ownership in the community; (b) data storage within jurisdictions the community controls or trusts; (c) explicit, informed, and ongoing consent for each use of the data, including training, evaluation, deployment, and publication; and (d) the unconditional right of the community to withdraw consent and require deletion at any time. Communities should be encouraged to develop their own data governance policies, and these policies should take precedence over the preferences of the research or technology organisation.

8.3 Benefit Distribution

Extractive models, where a technology or research organisation collects community data and retains all intellectual, commercial, or reputational benefit, are ethically unacceptable. They will also, in the medium term, undermine the trust that makes future data collection possible. Benefit distribution mechanisms should be agreed in advance of data collection and documented in a community data agreement. The Saving Voices Project makes all models and datasets freely available to the community and the global research community, with full attribution to community contributors.

8.4 Transparency and Failure Mode Documentation

All speech models deployed in indigenous communities must be accompanied by documentation that describes their limitations honestly and accessibly. This includes: the conditions under which performance degrades below acceptable thresholds; speaker groups for whom performance is systematically worse; and known error types (tonal, morphological, register-specific) that are unresolved. This documentation must be translated into the target language and shared with the community before deployment, not after.

9. Open Research Questions

This position paper argues for a direction of travel, not a solved problem. The most pressing open research questions, drawn in part from the Saving Voices Project's experience, are identified below.

RQ1	How can speech models be made robust to the extreme acoustic variability of elder speakers in languages where younger speakers are rare or unavailable for data collection? Standard speaker normalisation techniques assume a broad speaker distribution that does not exist for many endangered languages.
RQ2	What is the minimum viable dataset for a usable indigenous language ASR model, and how should 'usable' be defined by and for the community? Current frugal approaches suggest meaningful results are achievable below 3 hours; the floor needs to be systematically characterised.
RQ3	How should models handle code-switching between indigenous and dominant regional languages? This is the default mode of speech in many indigenous communities, yet no existing ASR architecture handles it reliably for low-resource language pairs.
RQ4	Can community-generated evaluation datasets produced without professional transcribers be sufficiently reliable for benchmarking? The Soliga project tested this and found reasonable reliability, but validation across more languages is needed (Ardila et al., 2020).
RQ5	How should TER be standardised across the diverse tonal systems of indigenous languages? Lexical tone, grammatical tone, and tone sandhi require different measurement approaches, and a community of practice is needed to develop shared standards.
RQ6	What governance and data sovereignty frameworks are most effective for indigenous language AI projects? The CARE Principles provide a foundation, but their operationalisation in specific cultural and legal contexts requires empirical investigation (Carroll et al., 2020).

10. Conclusion

Every language is an irreplaceable repository of human knowledge, cultural memory, and cognitive diversity. When an indigenous language falls silent, something unique and irretrievable is lost. Voice technology, applied thoughtfully and in genuine partnership with speaker communities, offers one of the most powerful tools available for documentation and revitalisation. But that technology must be redesigned, not merely adapted, to serve indigenous and other under-represented language communities.

This paper has argued that progress requires three coordinated technical interventions: byte- and morpheme-aware tokenisation that preserves the linguistic signal of polysynthetic and morphologically rich languages; self-supervised multilingual pre-training on typologically diverse corpora, combined with parameter-efficient adaptation strategies designed for the

extreme low-resource settings characteristic of endangered languages; and a multi-dimensional evaluation framework that replaces WER's false universality with metrics calibrated to tonal, morphological, and register-specific properties that actually matter for indigenous communities.

These technical interventions must be grounded in the CARE Principles for Indigenous Data Governance: community benefit, authority to control, responsibility, and ethics. The communities whose languages this agenda is designed to serve must be partners in every stage of the work, from orthographic design through model deployment and ongoing stewardship. Their knowledge, their priorities, and their voices must shape the research agenda, not merely serve as its data source.

The Saving Voices Project's work with the Soliga community of Karnataka provides a replicable blueprint. It demonstrates that meaningful voice documentation is achievable with frugal AI methods, community-champion data collection, and genuine respect for community data sovereignty. The goal is not to build the most sophisticated model; it is to build the most useful one, for the communities that need it most, before the window closes.

References

Ardila, R., Branson, M., Davis, K., Henretty, M., Kohler, M., Meyer, J., Morais, R., Saunders, L., Tyers, F.M. and Weber, G. (2020) 'Common Voice: A massively multilingual speech corpus', Proceedings of the 12th Language Resources and Evaluation Conference (LREC 2020), Marseille, pp. 4218-4222.

Baevski, A., Zhou, H., Mohamed, A. and Auli, M. (2020) 'wav2vec 2.0: A framework for self-supervised learning of speech representations', Advances in Neural Information Processing Systems (NeurIPS 2020), vol. 33, pp. 12449-12460.

Carroll, S.R., Garba, I., Figueroa-Rodriguez, O.L., Holbrook, J., Lovett, R., Materechera, S., Parsons, M., Raseroka, K., Rodriguez-Lonebear, D., Rowe, R., Sara, R., Walker, J.D., Anderson, J. and Hudson, M. (2020) 'The CARE Principles for Indigenous Data Governance', Data Science Journal, vol. 19, no. 1, p. 43.

Conneau, A., Baevski, A., Collobert, R., Mohamed, A. and Auli, M. (2021) 'Unsupervised cross-lingual representation learning for speech recognition', Proceedings of Interspeech 2021, Brno, pp. 2426-2430.

Everson, M. and Lillie, J. (eds.) (2001) The Unicode Standard, Version 3.0. Reading, MA: Addison-Wesley.

Fleurs, A., Conneau, A., Gangi, M.A., Ma, X. and Jegou, H. (2022) 'FLEURS: Few-shot learning evaluation of universal representations of speech', Proceedings of the 2022 IEEE Spoken Language Technology Workshop (SLT 2022), Doha, pp. 798-805.

Gut, U. (2008) 'Nigerian English: Phonology', in Schneider, E. (ed.) Varieties of English. Berlin: Mouton de Gruyter, pp. 35-54.

Hammarstrom, H., Forkel, R., Haspelmath, M. and Bank, S. (2023) Glottolog 5.0. Jena: Max Planck Institute for Evolutionary Anthropology. Available at: https://glottolog.org.

Himmelmann, N.P. (2006) 'Language documentation: What is it and what is it good for?', in Gippert, J., Himmelmann, N.P. and Mosel, U. (eds.) Essentials of Language Documentation. Berlin: Mouton de Gruyter, pp. 1-30.

Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L. and Chen, W. (2022) 'LoRA: Low-rank adaptation of large language models', Proceedings of the 10th International Conference on Learning Representations (ICLR 2022), Virtual.

Kamper, H., Goldwater, S., Livescu, K. and Shakhnarovich, G. (2022) 'Word segmentation on extremely low-resource languages using segmented language model', Proceedings of ICASSP 2022, Singapore, pp. 7962-7966.

Ludeling, A. and Kyto, M. (eds.) (2008) Corpus Linguistics: An International Handbook. Berlin: De Gruyter.

Muller, B., Ousidhoum, N., Omrani, M., Romanov, A., Cao, Y. and Sagot, B. (2021) 'Unseen languages are not unseen: A survey of zero-shot multilingual speech recognition', Findings of the Association for Computational Linguistics (ACL 2021), pp. 3249-3261.

Prabhavalkar, R., Hori, T., Sainath, T.N., Schluter, R. and Watanabe, S. (2023) 'End-to-end speech recognition: A survey', IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 31, pp. 2400-2430.

Radford, A., Kim, J.W., Xu, T., Brockman, G., McLeavey, C. and Sutskever, I. (2023) 'Robust speech recognition via large-scale weak supervision', Proceedings of the 40th International Conference on Machine Learning (ICML 2023), Honolulu, vol. 202, pp. 28492-28518.

Schuller, B., Steidl, S., Batliner, A., Burkhardt, F., Devillers, L., Muller, C. and Narayanan, S. (2013) 'The INTERSPEECH 2013 Computational Paralinguistics Challenge: Social signals, conflict, emotion, autism', Proceedings of INTERSPEECH 2013, Lyon, pp. 148-152.

Sennrich, R., Haddow, B. and Birch, A. (2016) 'Neural machine translation of rare words with subword units', Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL 2016), Berlin, pp. 1715-1725.

UNESCO (2003) Language Vitality and Endangerment. Paris: UNESCO Ad Hoc Expert Group on Endangered Languages.

Xue, L., Constant, N., Roberts, A., Kale, M., Al-Rfou, R., Siddhant, A., Barua, A. and Raffel, C. (2022) 'ByT5: Towards a token-free future with pre-trained byte-to-byte models', Transactions of the Association for Computational Linguistics, vol. 10, pp. 291-306.