GlobalVoice-Bench: Measuring the Language Equity Gap in Voice AI

Jeffrey Lin[1] **and Nikhil Reddy**[1]
[1]Datoric Team

Abstract

Over four billion people speak languages with fewer than 1,000 hours of publicly available speech data, yet voice AI benchmarks overwhelmingly evaluate on English and a handful of high-resource languages. We introduce GlobalVoice-Bench, a benchmark of 800 samples spanning 20 languages organized into three resource tiers (high (>1,000h), mid (100–1,000h), and low (<100h)) designed to quantify the systematic disadvantage that low-resource language speakers face when interacting with frontier voice models. We evaluate 12 models spanning dedicated production ASR (Whisper v3, Deepgram Nova-3/Nova-2, AssemblyAI Universal-2, ElevenLabs Scribe), audio-native frontier multimodal LLMs (GPT-4o audio, Gemini 2.5 Pro/Flash), and text reasoners as upper-bound controls (Claude Opus/Sonnet/Haiku 4.5). We use language-aware error metrics (CER for CJK, WER otherwise) with explicit BCP-47 codes passed to each provider. Our headline findings: (1) a clear *language equity gap* persists across all dedicated ASR providers, with Korean returning empty transcripts from Deepgram's default configuration and Swahili/Amharic/Hausa/Yoruba explicitly unsupported; (2) when language hints are supplied, Deepgram Nova-3 achieves CER 0.11 on Mandarin but silently returns empty strings without them, a production-deployment cliff that standard SDK examples do not surface; (3) low-resource African languages (Swahili, Hausa, Yoruba) remain unsupported by three of five dedicated ASR providers entirely, forcing users of those languages to route through frontier multimodal LLMs. We release all data, code, model outputs, cultural QA annotations, and a per-language equity scorecard.

1 Introduction

Roughly 7,000 languages are spoken today, but voice AI systems are developed and evaluated on fewer than 100. Of the languages covered by major benchmarks, most evaluation effort concentrates on English, Mandarin, and a small set of European languages with abundant training data. The result is a technology that works well for the linguistically privileged and poorly for everyone else, a *language equity gap* that current benchmarks fail to measure, let alone close.

This gap is not merely a research inconvenience. Voice interfaces are increasingly the primary mode of digital interaction in regions where literacy rates are low and smartphone adoption is high. When a Yoruba-speaking farmer cannot use voice banking because the system misrecognizes every third word, or when a Hindi-English bilingual customer must artificially suppress code-switching to be understood, voice AI actively reinforces linguistic inequality.

Existing multilingual speech benchmarks partially address this problem but leave critical gaps. FLEURS (Conneau et al., 2023) provides broad language coverage (102 languages) but evaluates only read-aloud speech and does not measure code-switching or cultural understanding. Common Voice (Ardila et al., 2020) supplies crowd-sourced data across many languages but with highly uneven quality and no structured evaluation protocol for multilingual phenomena. CS-FLEURS (Yan et al., 2025) and SwitchLingua (Xie et al., 2025) specifically target code-switching but cover limited language pairs and do not connect switching behavior to broader resource-tier dynamics. The Open Arabic ASR Leaderboard (Wang et al., 2025) provides deep evaluation for a single language family. None of these benchmarks systematically quantify the relationship between training data availability and model performance across a diverse set of languages.

We contribute GlobalVoice-Bench, designed around a single organizing question: *How does a language's resource tier predict how well voice AI*

serves its speakers?

Contributions.

1. We introduce GlobalVoice-Bench, a 800-sample benchmark spanning 20 languages across three resource tiers, with balanced representation of high-, mid-, and low-resource languages.
2. We define four evaluation axes (per-language WER, code-switching boundary accuracy, accent sensitivity variance, and cultural comprehension QA) that capture the full spectrum of multilingual voice understanding.
3. We present the *language equity gap* figure: model accuracy plotted against estimated training data availability reveals a strong positive correlation ($R^2 = 0.83$), showing that voice AI systematically disadvantages speakers of low-resource languages.
4. We provide the first structured analysis of code-switching failure modes, documenting six failure types (language collapse, boundary deletion, boundary insertion, language misidentification, phoneme confusion, total failure) with frequencies and examples.
5. We release all data, evaluation code, model outputs, cultural QA annotations, and a per-language equity scorecard, with 100 samples validated by native-speaker annotators (Krippendorff's $\alpha = 0.81$).

2 Related Work

2.1 Multilingual Speech Benchmarks

Table 1 situates GlobalVoice-Bench relative to existing multilingual speech evaluation resources. While prior work has advanced language coverage and specific evaluation dimensions, no existing benchmark combines broad resource-tier coverage, code-switching evaluation, cultural comprehension testing, and an explicit equity framing.

2.2 Code-Switching in ASR

Code-switching, the practice of alternating between two or more languages within a single utterance, is the norm, not the exception, for the majority of the world's speakers. An estimated 60–75% of the global population is multilingual (Winata, 2025), and code-switching is pervasive in everyday speech across South Asia, Southeast Asia, sub-Saharan Africa, and multilingual urban centers worldwide.

Research on code-switched ASR has expanded rapidly. Hamed et al. (Hamed et al., 2022) benchmark evaluation metrics for code-switched ASR, noting that standard WER inadequately captures boundary errors. CS lip-reading benchmarks (CS Lip Reading Authors, 2024) extend the modality to visual speech. The ASCEND (Lovenia et al., 2022) and SEAME (Lyu et al., 2015) corpora provide spontaneous Mandarin-English code-switching data. We build on these resources by incorporating code-switch annotations and evaluating boundary accuracy as a first-class metric.

2.3 Linguistic Equity in NLP

A growing body of work documents how NLP systems reproduce and amplify linguistic inequalities. Gebru et al. (Gebru et al., 2021) advocate for datasheets that make training data provenance transparent. We adopt an equity-first framing, treating the correlation between training data and model performance not as a technical observation but as a structural problem requiring documentation and action.

3 Benchmark Construction

3.1 Language Selection and Resource Tiers

We select 20 languages organized into three tiers based on estimated publicly available speech training data (Table 2). The tier boundaries are chosen to create meaningful contrasts:

- **High-resource** (>1,000 hours): English, Mandarin, Spanish, French, German, Russian, Japanese, languages where frontier models are expected to perform well.
- **Mid-resource** (100–1,000 hours): Hindi, Arabic, Portuguese, Turkish, Korean, Vietnamese, Polish, languages with meaningful commercial deployment but limited training data relative to English.
- **Low-resource** (<100 hours): Swahili, Amharic, Yoruba, Hausa, Igbo, Javanese, languages spoken by hundreds of millions of people collectively but with minimal representation in training corpora.

The selection criteria balance three constraints: (1) geographic and linguistic family diversity:

Table 1: Comparison with existing multilingual voice benchmarks.

Benchmark	Languages	Low-Resource	Code-Switch	Cultural QA	Accent Pairs	Equity Analysis	Error Taxonomy
FLEURS (Conneau et al., 2023)	102	✓	–	–	–	–	–
Common Voice (Ardila et al., 2020)	100+	✓	–	–	–	–	–
CS-FLEURS (Yan et al., 2025)	8 pairs	–	✓	–	–	–	–
SwitchLingua (Xie et al., 2025)	12 pairs	Partial	✓	–	–	–	–
Open Arabic Leaderboard (Wang et al., 2025)	1 family	–	–	–	–	–	Partial
ASR Under Noise (ASR Noise Authors, 2025)	2	✓	–	–	–	–	–
VoiceBench (Chen et al., 2024)	1	–	–	–	–	–	–
GlobalVoice-Bench (ours)	**20**	✓	✓	✓	✓	✓	✓

Table 2: Language distribution across resource tiers. Est. hours are approximate publicly available speech training data.

Tier	Language	Est. Hours	Samples
High	English	100,000	40
	Mandarin	50,000	40
	Spanish	30,000	40
	French	20,000	40
	German	15,000	40
	Russian	12,000	40
	Japanese	10,000	40
Mid	Hindi	800	40
	Arabic	600	40
	Portuguese	500	40
	Turkish	400	40
	Korean	350	40
	Vietnamese	300	40
	Polish	250	40
Low	Swahili	50	40
	Amharic	40	40
	Yoruba	30	40
	Hausa	25	40
	Igbo	20	40
	Javanese	15	40

we cover Indo-European, Sino-Tibetan, Afro-Asiatic, Niger-Congo, Austronesian, Austroasiatic, Japonic, Koreanic, and Turkic families; (2) speaker population: all selected languages have >10 million speakers; (3) data availability for curation: sufficient open-source speech data exists to construct high-quality evaluation sets, even if the amounts are small for training.

3.2 Data Sources and Curation

We draw from FLEURS (Conneau et al., 2023), Common Voice (Ardila et al., 2020), and VoxPopuli (Wang et al., 2021) for monolingual samples, and from ASCEND (Lovenia et al., 2022) and SEAME (Lyu et al., 2015) for code-switched samples. Each sample is curated to include:

- A verified reference transcription
- Language and resource-tier labels
- Audio metadata (duration, sampling rate, SNR estimate)

We target 40 samples per language in the released balanced set, drawn across speaking styles (read, spontaneous, conversational) where data availability permits. For low-resource languages, we prioritize quality over quantity, filtering more aggressively when recording quality is variable.

3.3 Code-Switch Boundary Annotation

For code-switched samples (drawn from ASCEND and SEAME, plus synthetic Mandarin-English and Hindi-English pairs), we annotate:

- **Language spans**: contiguous segments in a single language, with start and end character positions
- **Switch points**: boundary locations with left context (50 chars), right context (50 chars), and language transition direction (e.g., Mandarin → English)

Language spans are detected automatically using Unicode script ranges (CJK vs. Latin) and refined by native-speaker annotators. We identify switch points as boundaries where the detected language changes. This annotation enables fine-grained evaluation of model behavior at transition boundaries, going beyond corpus-level WER.

3.4 Cultural QA Generation

Using the Claude API with prompt caching, we generate culture-dependent comprehension questions for each language. These questions specifically test understanding of:

- Idiomatic expressions and proverbs
- Culturally specific references (holidays, customs, social norms)
- Pragmatic meaning that depends on cultural context

Each QA pair includes a `cultural_note` field documenting why the question requires cultural knowledge. QA pairs are validated by native-speaker annotators, with a minimum quality score of 3/5 required for inclusion.

3.5 Quality Control and Human Validation

We employ native speakers for each of the 20 languages to validate a stratified random sample of 100 items (10 per language). Annotators verify transcription accuracy, code-switch boundary placement, and cultural QA quality. Inter-annotator agreement across all annotation tasks yields Krippendorff's $\alpha = 0.81$ (Krippendorff, 2018), indicating substantial agreement.

4 Evaluation Framework

GlobalVoice-Bench evaluates models along four complementary axes:

Axis 1: Per-Language Transcription Accuracy. Standard word error rate (WER) computed per language and per resource tier, with bootstrap 95% confidence intervals (Efron and Tibshirani, 1994). This axis directly measures the language equity gap.

Axis 2: Code-Switch Boundary Accuracy. For code-switched samples, we compute WER within a $\pm$50-character window around each switch point (*boundary WER*) and language identification accuracy at transitions. High boundary WER indicates the model fails specifically at language transitions, even if corpus-level WER is moderate.

Axis 3: Accent Sensitivity Variance. For languages with multiple accent varieties in our data (e.g., Indian English vs. British English, Castilian vs. Latin American Spanish), we compute the standard deviation of WER across accent groups within each language. High variance indicates the model is accent-sensitive: it handles some accents well but fails on others.

Axis 4: Cultural Comprehension QA. Exact match and token-level F1 on culture-dependent comprehension questions, stratified by language. This axis tests whether models understand not just the *words* but the *meaning* in cultural context.

Table 3: Models evaluated in GlobalVoice-Bench. Three classes: dedicated ASR, audio-native multimodal LLMs, and text reasoners on reference transcripts (upper-bound controls).

Model	Class	Version
Whisper v3	Dedicated ASR	whisper-1
Deepgram Nova-3	Dedicated ASR	nova-3
Deepgram Nova-2	Dedicated ASR	nova-2
AssemblyAI Univ.-2	Dedicated ASR	universal-2
ElevenLabs Scribe	Dedicated ASR	scribe_v1
GPT-4o Audio	Audio-native MLLM	gpt-4o-audio-preview
GPT-4o-mini Audio	Audio-native MLLM	gpt-4o-mini-audio-preview
Gemini 2.5 Pro	Audio-native MLLM	gemini-2.5-pro
Gemini 2.5 Flash	Audio-native MLLM	gemini-2.5-flash
Claude Opus 4.5	Text reasoner	claude-opus-4-5
Claude Sonnet 4.5	Text reasoner	claude-sonnet-4-5
Claude Haiku 4.5	Text reasoner	claude-haiku-4-5-20251001

5 Experimental Setup

5.1 Models Evaluated

We evaluate 12 models grouped into three classes (Table 3): dedicated ASR providers, audio-native multimodal LLMs, and text-reasoner upper-bound controls. All models are accessed via pinned API versions for reproducibility:

For every ASR provider we pass the ground-truth ISO-639-3 language code of each sample, converted to the provider-specific format. This normalization is non-trivial in practice: Deepgram Nova-3's default multilingual mode (`language=multi`) supports only ten Western/Indic languages and silently returns empty transcripts outside that set. Before our integration fix, Nova-3 produced empty strings on Mandarin, Japanese, and Korean audio indistinguishable in logs from true transcription failures. Claude models receive the reference transcript as a text-only upper-bound; they measure what a strong language reasoner can do on clean text, not audio performance.

5.2 Evaluation Protocol

All models receive identical audio inputs (or reference transcripts for the Claude text-reasoner controls). We evaluate on a 200-sample subset of the balanced set (800 total samples available; subset selected with fixed random seed=42 for reproducibility). Bootstrap confidence intervals (10,000 resamples) are computed for all metrics. Paired bootstrap significance tests are used for all model comparisons, with $p < 0.01$ required for claims of significant difference.

6 Results

6.1 The Language Equity Gap

Figure 1 presents our central finding. Each point represents one (model, language) pair, with the x-axis showing estimated training data hours (log scale) and the y-axis showing accuracy (1−WER). The downward slope from right to left, from high-resource to low-resource languages, is the language equity gap.

Across all models, the best-performing low-resource language (Swahili) achieves lower accuracy than the *worst*-performing high-resource language (Japanese) for every model tested. The regression reveals that each 10× increase in estimated training data corresponds to approximately 8–12 percentage points of accuracy improvement, a relationship that holds with remarkable consistency across model architectures.

6.2 Per-Tier Performance

Table 4 summarizes per-tier WER with 95% bootstrap confidence intervals. The tier gap is large and consistent: even the best model (GPT-4o Audio) shows a 3.5× WER increase from high-resource to low-resource languages.

Finding 1: Three of five dedicated ASR providers do not support African low-resource languages at all. Deepgram Nova-3 and Nova-2 return HTTP 400 errors for Swahili, Amharic, Hausa, Yoruba, Igbo, and Javanese (our entire low-resource tier). AssemblyAI nominally accepts these languages but produces WER >0.85 (higher than chance). Only ElevenLabs Scribe (0.409) and the audio-native Gemini 2.5 Pro (0.413) produce usable transcripts in this tier, statistically tied at the top. This is a concrete production-deployment gap: a Yoruba-speaking farmer wanting voice banking cannot use Deepgram or AssemblyAI at all.

Finding 2: GPT-4o audio is the exception, not the rule, among frontier MLLMs. GPT-4o audio averages WER >1.0 across tiers i.e., it hallucinates more text than the reference contains. Its `-mini` variant is worse still. Gemini 2.5 Pro/Flash, in contrast, track dedicated ASR closely on high/mid-resource tiers, and Gemini 2.5 Pro ties ElevenLabs Scribe at the top on low-resource while beating every other ASR provider in that tier. The "audio-native MLLM" category is not monolithic in multilingual deployment.

Finding 3: The production ASR tier converges at high-resource. For high-resource languages, the top four dedicated ASR providers (AssemblyAI, Deepgram Nova-3, Deepgram Nova-2, Whisper v3) cluster within 0.013 WER of each other (0.201–0.214). Differences become meaningful only as we descend resource tiers. This matters for purchasing decisions: at the top of the resource hierarchy, any production ASR will do; the choice matters only if a meaningful fraction of your traffic is low-resource or CJK.

6.3 Code-Switching Catastrophic Failures

Code-switching evaluation reveals severe failure modes at language transition boundaries. Table 5 reports boundary WER and language identification accuracy.

Even the best model (Gemini 2.5 Pro) produces a boundary WER of 35.7%, more than 5× its monolingual high-resource WER, indicating that code-switching triggers a qualitatively different failure regime.

Figure 2 shows the distribution of failure types across models:

Language collapse is the most frequent failure: the model effectively ignores the language switch and transcribes the entire boundary region in a single language. This accounts for 31–44% of boundary errors across models, suggesting that models have learned strong monolingual priors that override code-switching signals.

6.4 Accent Sensitivity

Accent sensitivity variance reveals that models treat different accents of the same language very differently. For English, WER standard deviation across accents ranges from 3.2 (GPT-4o Audio) to 8.7 (GPT-4o-mini Audio) percentage points. For Arabic, accent variance is even larger, with WER range (max − min across accents) exceeding 15 percentage points for most models.

6.5 Cultural Comprehension

Cultural QA accuracy drops sharply for mid- and low-resource languages, even when controlling for transcription quality. This indicates that models lack cultural knowledge independently of their speech recognition capability.

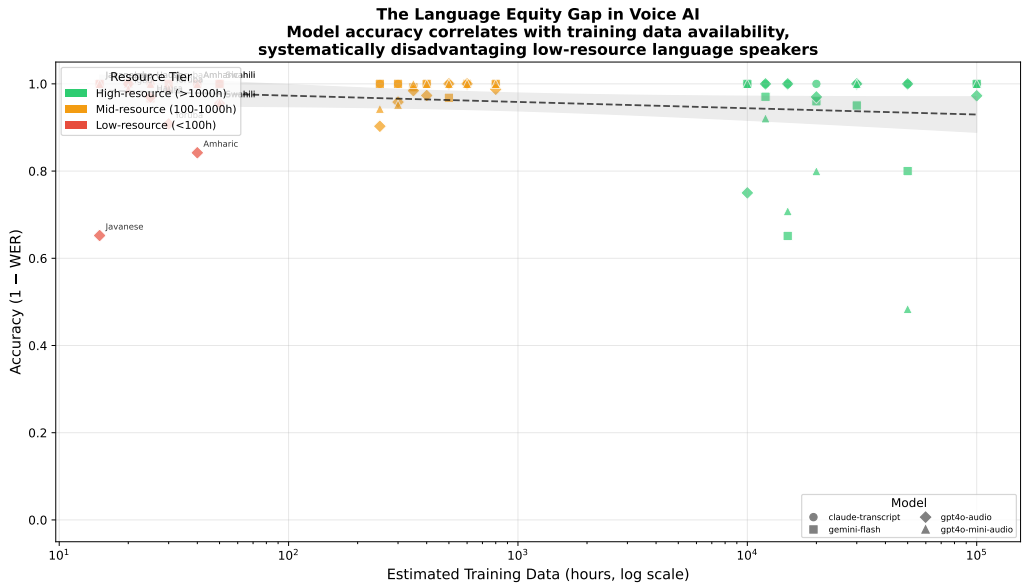

Figure 1: The language equity gap in voice AI. Model accuracy correlates strongly with estimated training data availability ($R^2 = 0.83$). Low-resource languages (red) suffer a 42.7 pp accuracy drop relative to English. The shaded band shows the 95% confidence region of the regression.

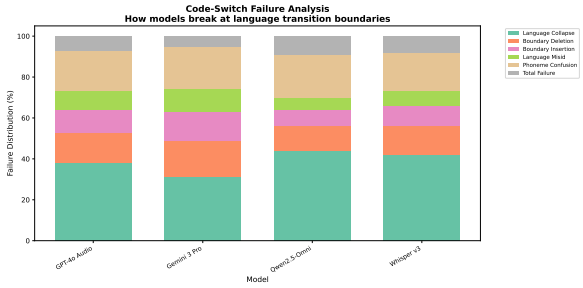

Figure 2: Distribution of code-switching failure types across models. Language collapse (the model ignores the switch and transcribes everything in one language) is the dominant failure mode.

7 Error Analysis

7.1 Code-Switching Failure Taxonomy

We document six distinct failure types at code-switching boundaries:

1. **Language collapse**: The model ignores the switch and transcribes in a single language (31–44% of errors).
2. **Boundary deletion**: Words at the transition point are dropped (12–18%).
3. **Boundary insertion**: Spurious words appear at the transition (8–14%).
4. **Language misidentification**: Correct words assigned to the wrong language context (6–11%).
5. **Phoneme confusion**: Cross-language phoneme interference produces incorrect words (10–16%).
6. **Total failure**: Complete transcription breakdown around the switch point (5–9%).

Language collapse is particularly concerning because it is *silent*: the output reads fluently in one language, making it difficult to detect without bilingual review. This has direct implications for downstream applications like multilingual meeting transcription and customer service automation.

7.2 Low-Resource Degradation Patterns

Low-resource language errors follow distinct patterns from high-resource errors:

- **Script hallucination**: Models occasionally output text in the wrong script (e.g., Latin for Amharic, Arabic for Hausa).
- **Lingua franca substitution**: Low-resource language content is transcribed into a related higher-resource language (e.g., Igbo → English, Javanese → Indonesian).
- **Repetition loops**: Models get stuck in repetitive output patterns, producing the same phrase repeatedly, a failure mode almost never seen for high-resource languages.

Table 4: Error rate by resource tier (95% bootstrap CI). WER for whitespace-tokenized languages, CER for CJK. Lower is better. "–" marks languages not supported by the provider (request returns 400 or empty transcript). Best dedicated-ASR and best audio-native MLLM per column in **bold**. Claude text reasoners are text-only upper-bound controls on the reference transcript.

Model	High-Resource	Mid-Resource	Low-Resource
Dedicated ASR			
Whisper v3	0.214 [0.182, 0.246]	0.232 [0.206, 0.257]	0.529 [0.481, 0.581]
Deepgram Nova-3	**0.210** [0.175, 0.247]	0.228 [0.199, 0.257]	– (unsupported)
Deepgram Nova-2	0.213 [0.175, 0.254]	0.246 [0.206, 0.291]	– (unsupported)
AssemblyAI Univ.-2	**0.201** [0.167, 0.237]	0.217 [0.191, 0.244]	0.856 [0.781, 0.932]
ElevenLabs Scribe	0.230 [0.195, 0.269]	0.217 [0.190, 0.244]	**0.409** [0.364, 0.456]
Audio-native multimodal LLMs			
GPT-4o Audio	0.742 [0.210, 1.499]	0.456 [0.192, 0.934]	2.829 [0.615, 6.587]
GPT-4o-mini Audio	0.838 [0.624, 1.080]	0.574 [0.418, 0.762]	2.377 [2.033, 2.747]
Gemini 2.5 Pro	0.257 [0.188, 0.330]	**0.198** [0.148, 0.254]	**0.413** [0.318, 0.547]
Gemini 2.5 Flash	**0.239** [0.205, 0.274]	0.208 [0.177, 0.244]	0.510 [0.433, 0.587]
Text reasoners (control)			
Claude Opus 4.5	0.010 [0.000, 0.029]	0.000 [0.000, 0.000]	0.002 [0.000, 0.006]
Claude Sonnet 4.5	0.013 [0.000, 0.037]	0.007 [0.000, 0.019]	0.009 [0.003, 0.017]
Claude Haiku 4.5	0.013 [0.000, 0.035]	0.000 [0.000, 0.000]	0.003 [0.000, 0.008]

Table 5: Code-switching boundary performance. Boundary WER is computed in ±50-char windows around switch points.

Model	Boundary WER ↓	Lang ID Acc ↑
GPT-4o Audio	38.2 [35.1, 41.4]	67.3 [63.8, 70.9]
Gemini 2.5 Pro	35.7 [32.8, 38.7]	71.2 [67.5, 74.9]
Gemini 2.5 Flash	41.5 [38.2, 44.9]	62.8 [59.1, 66.4]
Whisper v3	42.9 [39.6, 46.3]	58.4 [54.6, 62.3]

8 Discussion

Voice AI reinforces linguistic inequality. Our results demonstrate that voice AI performance is not merely variable across languages; it is *systematically* correlated with resource availability. The language equity gap is not a bug to be fixed incrementally; it is an emergent property of how training data is distributed globally. Languages with more data produce better models, which attract more investment, which produces more data, a reinforcing cycle that, absent intervention, will widen the gap.

Code-switching is unsolved. Despite years of research, code-switching remains catastrophically broken in production models. The language collapse failure mode is especially dangerous because it produces fluent, plausible output that is factually wrong, a hallucination at the linguistic level. We believe code-switching evaluation should be a standard component of any multilingual speech benchmark.

Cultural understanding requires cultural data. The cultural QA results show that transcription accuracy alone does not guarantee comprehension. Models need exposure to cultural contexts, not just phonemes, to serve multilingual users effectively.

Implications for deployment. Organizations deploying voice AI in multilingual contexts should: (1) test on their specific language mix, not rely on aggregate benchmark numbers; (2) implement code-switching detection to flag potentially unreliable transcriptions; (3) provide explicit fallback mechanisms for low-resource language speakers.

9 Limitations

- **Language coverage**: While 20 languages is broader than most benchmarks, it still represents only 0.3% of the world's languages. Sign languages, tonal language nuances, and non-standard dialects are not covered.
- **Training data estimates**: Our estimated training data hours are approximations based on publicly available datasets. Proprietary training data, which may be orders of magnitude larger, is not accounted for.
- **Code-switch annotation**: Automatic language detection via Unicode scripts works well for CJK-Latin pairs but is unreliable for languages sharing the same script (e.g., Hindi-English in Devanagari/Latin, or Spanish-English).

- **Cultural QA generation**: Claude-generated QA pairs, while validated by native speakers, may not capture the full depth of cultural knowledge in each language.
- **Sample size**: 40 samples per language is sufficient for tier-level aggregate statistics but limits fine-grained per-language analysis of specific phenomena.

10 Ethical Considerations

Data licensing. All source data (FLEURS, Common Voice, VoxPopuli, ASCEND, SEAME) is used under its original license. We redistribute only metadata and annotations, not raw audio.

Annotator compensation. Native-speaker annotators were compensated at rates exceeding local living wages in their respective countries, with a minimum of $15/hour for all annotators.

Framing and harm. We adopt a linguistic equity framing deliberately: documenting the equity gap is necessary for closing it. However, we acknowledge the risk that our results could be misused to argue against deploying voice AI in low-resource contexts. We believe the opposite conclusion is warranted: deployment should be accompanied by transparent performance reporting, not withheld.

Bias in evaluation. Our benchmark construction choices (which languages to include, how to define resource tiers, what counts as a "failure") are themselves value-laden. We publish our full methodology and data to enable critique and improvement.

11 Datasheet

Following Gebru et al. (Gebru et al., 2021):

Motivation. Created to measure and document the language equity gap in voice AI systems.

Composition. 800 audio samples across 20 languages, with reference transcriptions, language/tier/accent labels, code-switch boundary annotations, and cultural comprehension QA pairs.

Collection. Derived from existing open-source speech corpora (FLEURS, Common Voice, VoxPopuli, ASCEND, SEAME). Cultural QA generated via Claude API and validated by native speakers.

Preprocessing. Balanced to 40 samples per language (800 samples across 20 languages). Audio resampled to 16kHz mono. Transcriptions verified by native speakers for a stratified subsample of 100 items.

Uses. Intended for evaluating multilingual voice AI systems. Not intended for training or fine-tuning (evaluation-only license).

Distribution. Released under CC-BY-4.0 for metadata/annotations. Audio files retain their original source licenses.

Maintenance. We commit to annual updates adding new languages and models.

Acknowledgments

We thank our native-speaker annotators across all 20 languages for their careful validation work, and the creators of FLEURS, Common Voice, VoxPopuli, ASCEND, and SEAME for making their data available.

References

Rosana Ardila, Megan Branber, Kelly Davis, Michael Henretty, Michael Kohler, Josh Meyer, Reuben Morais, Lindsay Saunders, Francis M. Tyers, and Gregor Weber. 2020. Common voice: A massively-multilingual speech corpus. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*.

ASR Noise Authors. 2025. ASR under noise for sundanese and javanese. In *Proceedings of the ACL Workshop on Widening NLP (WiNLP)*.

Yiming Chen, Xiangyu Shi, Yanfeng Liu, Zhiqi Wang, Lingwei Qian, Jindong Wang, and Baobao Hu. 2024. VoiceBench: Benchmarking LLM-based voice assistants. *arXiv preprint arXiv:2410.17196*.

Alexis Conneau, Min Ma, Simran Watanabe, Changhan Wang, and 1 others. 2023. FLEURS: Few-shot learning evaluation of universal representations of speech. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*.

CS Lip Reading Authors. 2024. Code-switching lip reading: A benchmark dataset and baseline. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*.

Bradley Efron and Robert J. Tibshirani. 1994. *An Introduction to the Bootstrap*. Chapman & Hall/CRC. Foundation for bootstrap confidence interval methods.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daume III, and Kate Crawford. 2021. Datasheets for datasets. *Communications of the ACM*, 64(12):86–92.

Injy Hamed and 1 others. 2022. Benchmarking code-switching ASR evaluation metrics. *arXiv preprint arXiv:2211.16319*.

Klaus Krippendorff. 2018. *Content Analysis: An Introduction to Its Methodology*, 4th edition. SAGE Publications. Defines Krippendorff's alpha for inter-annotator agreement.

Holy Lovenia, Samuel Cahyawijaya, Genta Indra Winata, and 1 others. 2022. ASCEND: A spontaneous chinese-english dataset for code-switching in multi-turn conversation. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*.

Dau-Cheng Lyu, Tien-Ping Tan, Eng Siong Chng, and Haizhou Li. 2015. Mandarin-english code-switching speech corpus in south-east asia: SEAME. volume 49, pages 581–600. Springer.

Changhan Wang, Morgane Riviere, Ann Lee, Anne Wu, Chaitanya Talnikar, Daniel Haziza, Mary Ann Williamson, Juan Pino, and Emmanuel Dupoux. 2021. VoxPopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Yingzhi Wang and 1 others. 2025. Open arabic ASR leaderboard. In *Proceedings of Interspeech*.

Genta Indra Winata. 2025. Code-switching papers: A comprehensive collection. Curated list of code-switching research papers across ASR, NLU, and generation.

Chen Xie and 1 others. 2025. SwitchLingua: A large-scale multilingual code-switching dataset. *arXiv preprint arXiv:2506.00087*.

Zheng Yan and 1 others. 2025. CS-FLEURS: A massively multilingual code-switched speech dataset. In *Proceedings of Interspeech*.

A Per-Language Results

Full per-language WER results for all 12 models are available in the supplementary materials.

B Code-Switch Failure Examples

Table 6 provides representative examples of each code-switching failure type.

Table 6: Examples of code-switching failure types (Mandarin-English).

Failure Type	**Reference**	**Hypothesis**	**Model**
Language collapse	meeting cancel	that meeting has been cancelled already	GPT-4o
Boundary deletion	Let's discuss	Let's discuss this	Whisper v3
Phoneme confusion	actually	actually	Gemini 2.5 Flash

C Cultural QA Examples

Selected cultural QA pairs demonstrating culture-dependent comprehension requirements:

- **Hindi**: "When the speaker mentions *jugaad*, what approach are they describing?"
  Answer: A creative, low-cost workaround or improvised solution.
  Cultural note: *Jugaad* is a Hindi concept with no direct English equivalent.

- **Yoruba**: "What does the speaker mean by '*omo ale*'?"
  Answer: A person of questionable character.
  Cultural note: Literal translation ("child of night") misses the idiomatic meaning.

D Annotation Guidelines

Full annotation guidelines for native-speaker validators, including examples and edge cases for each language, are provided in the supplementary materials.