# AfriSwitch: Evaluating and Adapting Multilingual ASR for In-the-Wild African Code-Switched Conversations

**Anonymous ACL submission**

## Abstract

Code-switching is pervasive in bilingual African conversations, yet most ASR systems assume monolingual input and are evaluated on curated benchmarks, producing brittle performance in real deployments. We present **AfriSwitch**, an empirical study combining: (i) a novel 20.40-hour human-transcribed in-the-wild code-switched benchmark spanning five African languages (Swahili, Kinyarwanda, Zulu, Amharic, Hausa) with switch-level language annotations, (ii) zero-shot benchmarking of seven open and closed ASR systems revealing large WER gaps versus monolingual benchmarks, and (iii) controlled ablations over initialization, synthetic data mixing, and parameter-efficient adaptation for Swahili-English. Code-Mixing Index (CMI) and Switching frequency (S*) analysis reveal previously unreported variations in code-switching behaviour across African languages. Our adaptation findings reveal that synthetic-to-real transfer is poor: models adapted on synthetic code-switch data improve substantially on synthetic test sets but show limited gains on natural conversational speech. Using a mix of natural monolingual and synthetic code-switched training corpora, we find that full fine-tuning matches or exceeds lightweight adapters while better preserving monolingual performance, likely a reflection of task complexity and domain drift from pretrained models, providing actionable guidance for adapting models for real-world multilingual ASR in Africa. PIER analysis enabled by AfriSwitch's switch-level tags exposes systematic suppression of embedded English words invisible to standard WER, positioning AfriSwitch as a more realistic evaluation framework for African code-switched ASR.

## 1 Introduction

Recent years have seen a rapid increase in publicly available speech resources for African languages, including AfriSpeech-200 (Olatunji et al., 2023), Google Waxal (Anonymous, 2026), NCHLT (Barnard et al., 2014), ALFFA (Gelas et al., 2012; Abate et al., 2005), Mozilla Common Voice (Ardila et al., 2020), Africa Next Voices (Marivate et al., 2025; Umuganda, 2025), and Naija Voices (Emezue et al., 2025). Complementing these, evaluation benchmarks such as AfriVox (Awobade et al., 2025), FLEURS (Conneau et al., 2022), SimbaBench (Elmadany et al., 2025), and AfriSpeech-MultiBench (Ashungafac et al., 2025) now enable systematic ASR assessment across languages and domains. Yet most benchmarks assume linguistic consistency within utterances, which does not reflect everyday multilingual African communication, where speakers freely alternate between English or French, lingua francas, and local languages.

While multilingual ASR systems have demonstrated improved generalization across languages (Omnilingual-ASR-Team et al., 2025; Radford et al., 2022; Pratap et al., 2023), their robustness to naturally occurring conversational code-switching remains understudied. FLEURS-CS (Yan et al., 2025) constructs code-switched utterances from read speech via synthetic mixing, and the South African Soap Opera corpus (Yılmaz et al., 2018) uses scripted broadcast dialogue; neither captures the spontaneous, context-driven switching of real conversation. Critically, both lack switch-level language annotations, preventing token-level evaluation of embedded-language recognition.

Synthetic data-mixing has been shown to improve code-switching ASR in low-resource settings (Babatunde et al., 2025; Dhawan et al., 2023; Pratapa et al., 2018), but often at the cost of catastrophic forgetting on monolingual speech (Babatunde et al., 2025). Prior architectural approaches—language prompt fusion (Yang et al., 2023), attention-guided adaptation (Aditya et al., 2024), encoder refining with language-aware decoding (Zhao et al., 2025)—have shown varied improvement, but none have been systematically evaluated on naturally occurring African code-switched speech.

In this work, we provide 20.40 hours of in-the-wild code-switched speech across five African languages, benchmark seven ASR systems under realistic conditions, and conduct controlled Swahili–English adaptation ablations with Point-of-Interest Error Rate (PIER) (Ugan et al., 2025) analysis to diagnose embedded-language recognition failures.

**Contributions.**

- **AfriSwitch benchmark**: 20.40 hours of human-transcribed conversational code-switched speech across five African languages with switch-level annotations.
- **Zero-shot evaluation**: Seven open and commercial ASR systems benchmarked, revealing large WER gaps versus monolingual benchmarks.
- **Adaptation study**: Controlled Swahili–English ablations quantifying the synthetic-to-real gap and PIER-based analysis of embedded English recognition failures.

## 2 AfriSwitch Dataset

### 2.1 Language Selection and Data Sources

To span multiple African regions, we select **Hausa** (West Africa), **Swahili, Kinyarwanda, Amharic** (East Africa), and **Zulu** (Southern Africa). Following Li et al. (2023), audio was drawn from publicly available YouTube videos and podcasts with permissive licenses, selected by bilingual annotators on African crowdsourcing platforms (Olatunji et al., 2023) for the presence of code-switching. The dataset will be fully open-sourced after the anonymity period under a Creative Commons Attribution 4.0 International (CC BY 4.0) license.

### 2.2 Processing and Annotation

Audio was segmented using VAD; shorter segments were concatenated up to 40 seconds to increase the probability of capturing code-switches, following Koluguri et al. (2025). Bilingual native speakers (college-educated, ages 18–35, paid $10–$50/hour) produced verbatim transcriptions with language-boundary tags (<en>, <sw>, etc.).

A three-stage quality pipeline was applied: (1) initial transcription and review; (2) independent meta-review by a second annotator; (3) automated language identification (LID) verification using Meta's fastText (Joulin et al., 2017) and NVIDIA NeMo (Kuchaiev et al., 2019), retaining only utterances with >90% LID confidence in both languages and confirmed switching therefore retaining only utterance with high quality human and model agreement.

### 2.3 Dataset Statistics

Table 1 summarizes the resulting corpus: 20.40 hours across 2,468 utterances with 17,292 code-switch events. To characterize the degree of mixing, we report two complementary metrics for each language variety.

**Switch points (S*).** Following Gambäck and Das (2016), a switch point is defined as any token preceded by a token with a different language tag. We count the number of such alternation points per utterance and report both the per-utterance average and corpus total.

**Code-Mixing Index (CMI).** We adopt the CMI formulation of Das and Gambäck (2014):

$$\mathrm{CMI} = \begin{cases} 100 \times \left[1 - \dfrac{\max(w_i)}{n-u}\right] & n > u \\ 0 & n = u \end{cases} \quad (1)$$

where $w_i$ is the token count for language $i$, $\max\{w_i\}$ is the count for the most prominent language, $n$ is the total number of tokens, and $u$ is the number of language-independent tokens. A CMI of 0 indicates a monolingual utterance; higher values reflect a more balanced distribution across languages (Srivastava and Singh, 2021).

### 2.4 All Code-Switching is NOT Equal

CMI scores range from 28.47 (Swahili) to 7.03 (Hausa), and S* from 9.81 (Kinyarwanda) to 4.89 (Hausa), reflecting genuine differences in mixing density, average number of language switches per utterance, and differing annotation granularity across languages. With the fewest switches and lowest CMI, Hausa is closer to monolingual. Swahili and Kinyarwanda, by contrast, with significantly higher CMI and S*, have a higher switch frequency with a more balanced representation of English.

**Training Datasets** Following

Table 1: AfriSwitch corpus statistics. **Avg S*** = average number of switch points per utterance (Gambäck and Das, 2016). **CMI** = Code-Mixing Index (Das and Gambäck, 2014).

| Lang | Hrs | Utt | Avg S* | Total S* | CMI |
|---|---|---|---|---|---|
| Swahili | 3.41 | 412 | 6.72 | 2,769 | 28.47 |
| Kinyarwanda | 3.84 | 464 | 9.81 | 4,554 | 27.70 |
| Zulu | 5.57 | 670 | 5.80 | 3,887 | 22.16 |
| Amharic | 5.40 | 649 | 7.31 | 4,746 | 15.11 |
| Hausa | 2.18 | 273 | 4.89 | 1,336 | 7.03 |
| **Total** | **20.40** | **2,468** | **7.01** | **17,292** | **20.73** |

Table 2: Zero-shot WER (%) on AfriSwitch. SW=Swahili, RW=Kinyarwanda, AMH=Amharic, ZU=Zulu, HA=Hausa. "–" = language not supported.

| Model | SW | RW | AMH | ZU | HA |
|---|---|---|---|---|---|
| Whisper-v3-Lg | 88.98 | 107.21 | 86.42 | 76.57 | 84.60 |
| XLS-R 2B | 112.34 | 129.78 | 104.56 | 95.42 | 110.83 |
| SeamlessM4T | 63.17 | 71.45 | 68.92 | 59.34 | 74.28 |
| OmniASR | 38.27 | **35.41** | **37.88** | 36.77 | 63.38 |
| Sahara | **26.82** | 35.48 | 52.13 | 32.23 | **30.46** |
| Gemini-3-flash | 31.68 | 58.10 | 44.71 | 24.31 | 42.61 |
| Azure Speech | 29.31 | – | 46.28 | **22.41** | – |

## 3 Benchmarking Multilingual ASR Systems

### 3.1 Evaluated Models and Protocol

We evaluate **Whisper-v3-Large** (Radford et al., 2022) (large encoder-decoder, weak-supervision training), **XLS-R 2B** (Babu et al., 2021) (self-supervised cross-lingual), **SeamlessM4T-v2-Large** (Seamless Communication, 2023) (speech recognition and translation), **OmniASR** (Omnilingual-ASR-Team et al., 2025) (1,600+ language coverage), **Sahara** (Africa-optimized production system), **Gemini-3-flash** (multimodal, commercial) and **Azure Speech** (commercial). This set spans open-source research models, large-scale multilingual models, production-API systems, and Africa-focused systems.

All models are evaluated zero-shot using default inference configurations. Language boundary tags are removed from references prior to scoring. Transcriptions are lowercased and punctuation-stripped before computing WER. No language prompts or tokens are provided, as real deployments rarely have per-utterance language labels.

### 3.2 Results and Analysis

Table 2 shows that *all* systems exhibit substantially higher WER than on standard monolingual benchmarks (Awobade et al., 2025; Elmadany et al., 2025), confirming that naturally occurring conversational code-switching is a significant failure mode.

Besides monolingual baseline performance, we observe a strong positive relationship between code-switch intensity (Avg $S^*$) and language-level zero-shot WER (Pearson $r = 0.79$, Spearman $\rho = 0.70$), suggesting that more frequent switching behavior (e.g. Kinyarwanda and Amharic) substantially increases recognition difficulty under zero-shot conditions.

Large multilingual models Whisper-v3-Large and XLS-R 2B show the worst performance (WER 76–130%), demonstrating that scale alone does not ensure robustness to bilingual speech. SeamlessM4T improves but still reaches 59–74% WER. The strongest results come from Africa-focused and high-coverage systems: Sahara (26–52% WER), Gemini-3-flash (24–58%), and Azure Speech (22–46%), suggesting that targeted language coverage and architecture choices matter more than parameter count. The performance gap between OmniASR/Sahara/Gemini and Whisper/XLS-R is consistent across all five languages, reinforcing that *multilingual pretraining alone is insufficient* for robust code-switch recognition.

## 4 Code-Switching Adaptation

### 4.1 Setup

We focus on **Swahili–English**, a major East African lingua franca with ≈92.2 million speakers (Ethnologue, 2026).

**Adaptation Evaluation Datasets** To complement the analysis on AfriSwitch, we evaluate finetuned ASR models on Swahili code-switched speech–CS-FLEURS (Yan et al., 2025)–monolingual Swahili speech from Afrivox-v1 (Awobade et al., 2025), and monolingual accented Engish utterances from Afrispeech-Multibench (Ashungafac et al., 2025).

**Adaptation Training Datasets** Monolingual training data (1,000 hrs English, 1,000 hrs Swahili) is drawn from AfriSpeech-200 (Olatunji et al., 2023), Mozilla Common Voice (Ardila et al., 2020), and African Next Voices (Umuganda, 2025).

**Synthetic code-switch generation.** Following Dhawan et al. (2023), we perform inter-sentential concatenation of monolingual utterances with two controls: (1) **duration-bucketed sampling** (1–5 s, 5–10 s, 10–20 s) to diversify switch-span distribution; (2) a **language-alternation constraint** ensuring consecutive segments come from different languages. This yields ≈3,000 hrs of synthetic Sw–En speech, forming a 5k-hour corpus (EN 1k + SW 1k + Synth 3k). An out-of-domain synthetic test set is constructed by mixing Swahili from AfriVox (Awobade et al., 2025) with English from AfriSpeech-MultiBench (Ashungafac et al., 2025).

**Models and fine-tuning.** We compare two architectures: **STT Conformer** (Rekesh et al., 2023) (120M params, initialized from either a monolingual English or Swahili checkpoint) and **Whisper-Small** (Radford et al., 2022) (244M, multilingual init). For each, we compare **full fine-tuning** (all parameters) versus **adapter-based tuning** (Houlsby et al., 2019) (backbone frozen, lightweight bottleneck layers in attention blocks). The STT Conformer uses an aggregated multilingual tokenizer (Dhawan et al., 2023; Sekoyan et al., 2025). All models train to convergence (≈80K steps) with AdamW and cosine schedule.

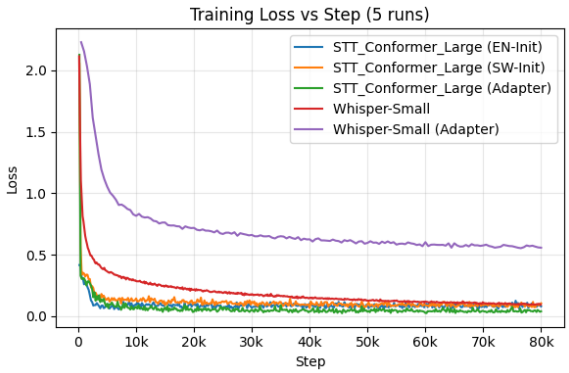


Figure 1: Training loss curves for the different fine-tuning strategies. Full FT reaches lower final loss than adapter tuning; Whisper adapter converges to a higher plateau, indicating limited capacity for parameter-efficient adaptation in that architecture.

## 4.2 Results

### 4.2.1 Finetuning Results

Table 3 presents WER across four test conditions evaluating three fine-tuning strategies—base (no adaptation), adapter tuning, and full fine-tuning—on a synthetic code-switch test set, the natural AfriSwitch corpus, the CS-FLEURS Swahili–English benchmark (Yan et al., 2025), and monolingual Swahili and English held-out sets (Awobade et al., 2025; Ashungafac et al., 2025).

Table 3: WER (%) under different adaptation strategies. EN-init STT adapter results are omitted as full fine-tuning consistently dominated for that initialisation.

| Model | Base | Adapter | Full FT |
|---|---|---|---|
| *Synthetic code-switch test set* | | | |
| Whisper-Small | 82.84 | 28.13 | **8.71** |
| STT SW-init | 58.89 | 34.96 | **18.42** |
| STT EN-init | 70.35 | — | **29.54** |
| *AfriSwitch (natural, in-the-wild)* | | | |
| Whisper-Small | 95.39 | **60.97** | 61.60 |
| STT SW-init | 87.67 | 65.78 | **55.68** |
| STT EN-init | 83.64 | — | 59.91 |
| *CS-FLEURS Swahili–English (Yan et al., 2025)* | | | |
| Whisper-Small | 111.23 | 46.80 | **37.88** |
| STT SW-init | 41.81 | 40.46 | **34.87** |
| STT EN-init | 90.53 | — | **39.50** |
| *Monolingual Swahili (Awobade et al., 2025)* | | | |
| Whisper-Small | 116.23 | 30.91 | **13.48** |
| STT SW-init | 17.80 | **14.00** | 17.61 |
| STT EN-init | 99.90 | — | **28.28** |
| *Monolingual English (Ashungafac et al., 2025)* | | | |
| Whisper-Small | 35.93 | 20.58 | **14.81** |
| STT SW-init | 99.11 | 32.10 | **27.75** |
| STT EN-init | 38.18 | — | **35.72** |

### 4.2.2 Code-Switch-Focused Evaluation: PIER

Table 4: PIER breakdown (Ugan et al., 2025) for best full fine-tuning checkpoints on AfriSwitch. English words are the points of interest. POI = points of interest (English words); Hits = correctly transcribed POI words. Presented results are average for all utterances

| Model | PIER↓ | Ins↓ | Del↓ | Sub↓ | Hits↑ | POI |
|---|---|---|---|---|---|---|
| Whisper-Small | 58.81 | 0.70 | 4.49 | 4.61 | 8.86 | 17.96 |
| STT SW-init | **49.00** | 0.40 | **3.34** | 4.34 | **10.28** | 17.96 |
| STT EN-init | 62.99 | **0.17** | 6.23 | **4.26** | 7.46 | 17.96 |

Since WER on code-switched speech can mask degradation on embedded-language tokens by overweighting the dominant matrix language (Ugan et al., 2025), we additionally report PIER (Point-of-Interest Error Rate) (Ugan et al., 2025) (Table 4), which restricts evaluation to English words only, for the best full fine-tuning checkpoint of each initialisation.

## 4.3 Analysis of Results

**Synthetic-to-real gap.** Adaptation yields large WER reductions on synthetic data (Whisper-Small: 82.84 → 8.71) and CS-FLEURS

(111.23 → 37.88), yet gains transfer poorly to AfriSwitch (95.39 → 61.60). Synthetic mixing captures lexical alternation but not the acoustic variability and spontaneous dynamics of real conversational speech, confirming AfriSwitch as a harder and more realistic benchmark.

**Initialization and adaptation strategy.** As shown in Table 3, SW-init achieves the best AfriSwitch WER (55.68) and lowest PIER (49.00), confirming that locally dominant language pretraining provides a stronger inductive bias. Adapter tuning preserves monolingual Swahili better than full FT (14.00 vs. 17.61), however, it fails to adapt adequately to natural code-switching in AfriSwitch (65.78 vs FT at 55.58). Finetuned Whisper easily overfits to the training data domain, achieving single-digit WER on the in-domain synthetic test set, lowest WER on monolingual data, competitive on CS-FLEURS, but fails to generalize adequately to AfriSwitch. Overall, and counterintuitively, full-finetuning outperforms parameter efficient adaptation approaches likely because the code-switched data domain represents a significant distribution shift from the base model's learned distribution.

**PIER: embedded-language failures.** WER masks failures on embedded English by overweighting Swahili. PIER reveals that even the best model (SW-init, 49.00) misses roughly half of all English words, with deletions dominating, peaking at 6.23 for EN-init indicating systematic suppression at switch boundaries. This analysis is enabled by AfriSwitch's language-level tags, which CS-FLEURS lacks.

## 5 Conclusion

We presented AfriSwitch, a 20.40-hour human-transcribed in-the-wild code-switched benchmark across five African languages, and a controlled adaptation study for Swahili-English. Our central finding is a clear synthetic-to-real gap: models adapted on synthetic data improve on matched conditions and CS-FLEURS, but still struggle on natural conversational speech. PIER analysis enabled by AfriSwitch's switch-level tags exposes systematic suppression of embedded English invisible to WER, positioning AfriSwitch as a more realistic evaluation framework for African code-switched ASR. Future work will scale to more language pairs and explore richer synthetic mixing strategies to close this gap.

## Limitations

While AfriSwitch represents a meaningful step toward realistic evaluation of code-switched ASR for African languages, several limitations should be noted.

**Language and geographic coverage.** AfriSwitch spans five African languages across three regions, but the continent hosts over 2,000 languages. The selected languages, while diverse, skew toward relatively higher-resourced African languages with existing ASR infrastructure. Lower-resourced languages and francophone African code-switching (e.g., French–Wolof, French–Lingala) are not represented.

**Adaptation scope.** The controlled adaptation study is limited to Swahili–English, a single language pair. The synthetic-to-real gap, initialization effects, and adapter trade-offs observed here may not generalize to other pairs with different typological properties, script systems, or switching patterns. Extending ablations to Kinyarwanda–English or Hausa–English, for instance, remains future work.

**PIER cross-benchmark comparison.** PIER evaluation is restricted to AfriSwitch due to the absence of switch-level language tags in CS-FLEURS (Yan et al., 2025). This prevents direct comparison of embedded-language error rates across benchmarks, limiting the generalizability of our PIER findings.

## Ethical Considerations

**Annotator compensation and working conditions.** All transcription annotators were bilingual native speakers recruited through African crowdsourcing platforms and compensated at $10–$50 per hour, above local market rates. Annotators were informed of the nature of the task, worked flexible hours, and were not exposed to harmful or sensitive content.

**Data provenance.** Audio was sourced from publicly available YouTube videos and podcasts under permissive licenses. Annotators performed verbatim transcription only; no original recordings were made. The released dataset will contain transcriptions and processed audio segments only, without links to original sources, preventing direct tracing back to the original content creators.

**Annotator biases.** Transcription annotators were college-educated bilingual speakers aged 18–35.

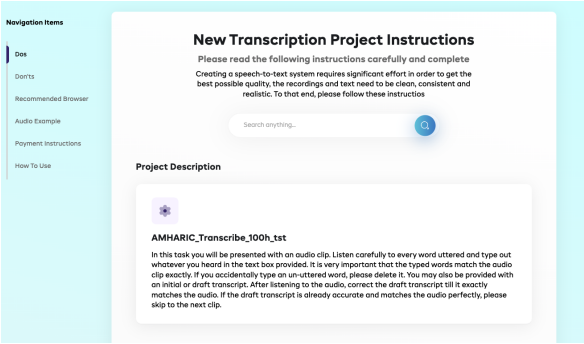


Figure 2: Annotator Instructions

While this ensures linguistic competence, it may introduce biases in how disfluencies, informal registers, and non-standard orthography are handled. No annotator demographic information is included in the released benchmark.

**Language representation.** By focusing on five African languages, this work implicitly prioritizes these communities over the many other African language communities not represented. We encourage the community to extend this benchmark to a broader set of languages, particularly those with fewer existing resources.

## References


Solomon Teferra Abate, Wolfgang Menzel, and Bairu Tafila. 2005. An amharic speech corpus for large vocabulary continuous speech recognition. In *INTERSPEECH-2005*.

Bobbi Aditya, Mahdin Rohmatillah, Liang-Hsuan Tai, and Jen-Tzung Chien. 2024. Attention-guided adaptation for code-switching speech recognition. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 10256–10260. IEEE.

Anonymous. 2026. Waxal: A large-scale multilingual african language speech corpus. *arXiv preprint arXiv:2602.02734*.

Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis Tyers, and Gregor Weber. 2020. Common voice: A massively-multilingual speech corpus. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4218–4222, Marseille, France. European Language Resources Association.

Gabrial Zencha Ashungafac, Mardhiyah Sanni, Busayo Awobade, Alex Gichamba, and Tobi Olatunji. 2025. AfriSpeech-MultiBench: A verticalized multidomain multicountry benchmark suite for African accented English ASR. In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pages 3642–3653, Mumbai, India. The Asian Federation of Natural Language Processing and The Association for Computational Linguistics.

Busayo Awobade, Mardhiyah Sanni, Tassallah Abdullahi, Chibuzor Okocha, Kelechi Ezema, Devendra Deepak Kayande, Lukman Enegi Ismaila, Tobi Olatunji, and Gloria Ashiya Katuka. 2025. Afrivox: Probing multilingual and accent robustness of speech llms. OpenReview / ACL ARR submission.

Oreoluwa Boluwatife Babatunde, Victor Tolulope Olufemi, Emmanuel Bolarinwa, Kausar Yetunde Moshood, and Chris Chinenye Emezue. 2025. Beyond monolingual limits: Fine-tuning monolingual ASR for Yoruba-English code-switching. In *Proceedings of the 7th Workshop on Computational Approaches to Linguistic Code-Switching*, pages 18–25, Albuquerque, New Mexico, USA. Association for Computational Linguistics.

Arun Babu, Changhan Wang, Andros Tjandra, Kushal Lakhotia, Qiantong Xu, Naman Goyal, Kritika Singh, Patrick von Platen, Yatharth Saraf, Juan Pino, Alexei Baevski, Alexis Conneau, and Michael Auli. 2021. Xls-r: Self-supervised cross-lingual speech representation learning at scale. *Preprint*, arXiv:2111.09296.

Etienne Barnard, Marelie Hattingh Davel, Charl Johannes van Heerden, Febe de Wet, and Jaco Badenhorst. 2014. The nchlt speech corpus of the south african languages. In *Workshop on Spoken Language Technologies for Under-resourced Languages*.

Alexis Conneau, Min Ma, Simran Khanuja, Yu Zhang, Vera Axelrod, Siddharth Dalmia, Jason Riesa, Clara Rivera, and Ankur Bapna. 2022. Fleurs: Few-shot learning evaluation of universal representations of speech. *arXiv preprint arXiv:2205.12446*.

Amitava Das and Björn Gambäck. 2014. Identifying languages at the word level in code-mixed Indian social media text. In *Proceedings of the 11th International Conference on Natural Language Processing*, pages 378–387.

Kunal Dhawan, KDimating Rekesh, and Boris Ginsburg. 2023. Unified model for code-switching speech recognition and language identification based on concatenated tokenizer. In *Proceedings of the 6th Workshop on Computational Approaches to Linguistic Code-Switching*, pages 74–82.

AbdelRahim A. Elmadany, Sang Yun Kwon, Hawau Olamide Toyin, Alcides Alcoba Inciarte, Hanan Aldarmaki, and Muhammad Abdul-Mageed. 2025. Voice of a continent: Mapping Africa's speech technology frontier. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 11028–11050, Suzhou, China. Association for Computational Linguistics.

Chris Emezue, NaijaVoices Community, Busayo Awobade, Abraham Owodunni, Handel Emezue, Gloria Monica Tobechukwu Emezue, Nefertiti Nneoma Emezue, Sewade Ogun, Bunmi Akinremi, David Ifeoluwa Adelani, and Chris Pal. 2025. The naijavoices dataset: Cultivating large-scale, high-quality, culturally-rich speech data for african languages. *Preprint*, arXiv:2505.20564.

Ethnologue. 2026. Swahili. Accessed: 2026-05-24.

Björn Gambäck and Amitava Das. 2016. Comparing the level of code-switching in corpora. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1850–1855.

Hadrien Gelas, Laurent Besacier, and Francois Pellegrino. 2012. Developments of Swahili resources for an automatic speech recognition system. In *SLTU - Workshop on Spoken Language Technologies for Under-Resourced Languages*, Cape-Town, Afrique Du Sud.

Neil Houlsby and 1 others. 2019. Parameter-efficient transfer learning for nlp. In *ICML*.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2017. Bag of tricks for efficient text classification. *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*.

Nithin Rao Koluguri, Monica Sekoyan, George Zelenfroynd, Sasha Meister, Shuoyang Ding, Sofia Kostandian, He Huang, Nikolay Karpov, Jagadeesh Balam, Vitaly Lavrukhin, and 1 others. 2025. Granary: Speech recognition and translation dataset in 25 european languages. *arXiv preprint arXiv:2505.13404*.

Oleksii Kuchaiev, Wei-Ning Hsu, Sahar Ghannay, and 1 others. 2019. Nemo: A toolkit for building ai applications using neural modules. In *NeurIPS Systems Demonstrations*.

Xinjian Li, Shinnosuke Takamichi, Takaaki Saeki, William Chen, Sayaka Shiota, and Shinji Watanabe. 2023. Yodas: Youtube-oriented dataset for audio and speech. In *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 1–8. IEEE.

Vukosi Marivate, Kayode Olaleye, Sitwala Mundia, Andinda Bakainga, Unarine Leo Netshifhefhe, Mahmooda Milanzie, Hope Mogale, THAPELO SINDANE, Zainab Abdulrasaq, Kesego Mokgosi, Chijioke Okorie, Nia Zion van Wyk, Graham Morrissey, Dale Dunbar, Francois Smit, Tsosheletso Chidi, Rooweither Mabuya, Andiswa Bukula, RESPECT MLAMBO, and 3 others. 2025. The south african next voices multilingual speech dataset.

Tobi Olatunji, Tejumade Afonja, Aditya Yadavalli, Chris Chinenye Emezue, Sahib Singh, Bonaventure F. P. Dossou, Joanne Osuchukwu, Salomey Osei, Atnafu Lambebo Tonja, Naome Etori, and Clinton Mbataku. 2023. AfriSpeech-200: Pan-African accented speech dataset for clinical and general domain ASR. *Transactions of the Association for Computational Linguistics*, 11:1669–1685.

Omnilingual-ASR-Team, Gil Keren, Artyom Kozhevnikov, Yen Meng, Christophe Ropers, Matthew Setzler, Skyler Wang, Ife Adebara, Michael Auli, Can Balioglu, Kevin Chan, Chierh Cheng, Joe Chuang, Caley Droof, Mark Duppenthaler, Paul-Ambroise Duquenne, Alexander Erben, Cynthia Gao, Gabriel Mejia Gonzalez, and 14 others. 2025. Omnilingual asr: Open-source multilingual speech recognition for 1600+ languages. *Preprint*, arXiv:2511.09690.

Vineel Pratap, Andros Tjandra, Bowen Shi, Paden Tomasello, Arun Babu, Sayani Kundu, Ali Elkahky, Zhaoheng Ni, Apoorv Vyas, Maryam Fazel-Zarandi, Alexei Baevski, Yossi Adi, Xiaohui Zhang, Wei-Ning Hsu, Alexis Conneau, and Michael Auli. 2023. Scaling speech technology to 1,000+ languages. *Preprint*, arXiv:2305.13516.

Adithya Pratapa, Gokul Bhat, Monojit Choudhury, Kalika Bali, and Dipti Sharma. 2018. Language modeling for code-mixing: The role of linguistic theory based synthetic data. In *ACL*.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision. *Preprint*, arXiv:2212.04356.

Dima Rekesh, Nithin Rao Koluguri, Samuel Kriman, Somshubra Majumdar, Vahid Noroozi, He Huang, Oleksii Hrinchuk, Krishna Puvvada, Ankur Kumar, Jagadeesh Balam, and 1 others. 2023. Fast conformer with linearly scalable attention for efficient speech recognition. In *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 1–8. IEEE.

Yu-An Chung Mariano Coria Meglioli David Dale Ning Dong Mark Duppenthaler Paul-Ambroise Duquenne Brian Ellis Hady Elsahar Justin Haaheim John Hoffman Min-Jae Hwang Hirofumi Inaguma Christopher Klaiber Ilia Kulikov Pengwei Li Daniel Licht Jean Maillard Ruslan Mavlyutov Alice Rakotoarison Kaushik Ram Sadagopan Abinesh Ramakrishnan Tuan Tran Guillaume Wenzek Yilin Yang Ethan Ye Ivan Evtimov Pierre Fernandez Cynthia Gao Prangthip Hansanti Elahe Kalbassi Amanda Kallet Artyom Kozhevnikov Gabriel Mejia Robin San Roman Christophe Touret Corinne Wong Carleigh Wood Bokai Yu Pierre Andrews Can Balioglu Peng-Jen Chen Marta R. Costa-jussà Maha Elbayad Hongyu Gong Francisco Guzmán Kevin Heffernan Somya Jain Justine Kao Ann Lee Xutai Ma Alex Mourachko Benjamin Peloquin Juan Pino Sravya Popuri Christophe Ropers Safiyyah Saleem Holger Schwenk Anna Sun Paden Tomasello Changhan Wang Jeff Wang Skyler Wang Mary Williamson Seamless Communication, Loïc Barrault. 2023. Seamless: Multilingual expressive and streaming speech translation.

Monica Sekoyan, Nithin Rao Koluguri, Nune Tadevosyan, Piotr Zelasko, Travis Bartley, Nikolay Karpov, Jagadeesh Balam, and Boris Ginsburg. 2025. Canary-1b-v2 parakeet-tdt-0.6b-v3: Efficient and high-performance models for multilingual asr and ast. *Preprint*, arXiv:2509.14128.

Vivek Srivastava and Mayank Singh. 2021. Challenges and limitations with the metrics measuring the complexity of code-mixed text. In *Proceedings of the Fifth Workshop on Computational Approaches to Linguistic Code-Switching*, pages 6–14.

Enes Yavuz Ugan, Ngoc-Quan Pham, Leonard Bärmann, and Alex Waibel. 2025. PIER: A novel metric for evaluating what matters in code-switching. In *ICASSP 2025 – 2025 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 1–5. IEEE.

Digital Umuganda. 2025. Afrivoice kinyarwanda speech dataset. https://huggingface.co/datasets/DigitalUmuganda/Afrivoice_Kinyarwanda.

Brian Yan, Injy Hamed, Shuichiro Shimizu, Vasista Lodagala, William Chen, Olga Iakovenko, Bashar Talafha, Amir Hussein, Alexander Polok, Kalvin Chang, Dominik Klement, Sara Althubaiti, Puyuan Peng, Matthew Wiesner, Thamar Solorio, Ahmed Ali, Sanjeev Khudanpur, Shinji Watanabe, Chih-Chen Chen, and 8 others. 2025. Cs-fleurs: A massively multilingual and code-switched speech dataset. *Preprint*, arXiv:2509.14161.

Yuhang Yang, Yizhou Peng, Xionghu Zhong, Hao Huang, and Eng Siong Chng. 2023. Adapting openai's whisper for speech recognition on code-switch mandarin-english seame and asru2019 datasets. *arXiv preprint arXiv:2311.17382*.

Emre Yılmaz, Astik Biswas, Ewald van der Westhuizen, Febe de Wet, and Thomas Niesler. 2018. Building a unified code-switching asr system for south african languages. *Preprint*, arXiv:1807.10949.

Jiahui Zhao, Hao Shi, Chenrui Cui, Tianrui Wang, Hexin Liu, Zhaoheng Ni, Lingxuan Ye, and Longbiao Wang. 2025. Adapting whisper for code-switching through encoding refining and language-aware decoding. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.