Dissecting ASR Failures in Low-Resource South Asian Languages

Abdul Hameed Azeemi, Ihsan Ayyub Qazi, Maryam Mustafa, Agha Ali Raza

Lahore University of Management Sciences, Pakistan

abdul.azeemi@lums.edu.pk, ihsan.qazi@lums.edu.pk, maryam_mustafa@lums.edu.pk, agha.ali.raza@lums.edu.pk

Abstract

Urdu, Punjabi, Pashto, and Sindhi, spoken by over 350 million people, pose unique ASR challenges: closely related Perso-Arabic scripts, unstandardized orthography, and mixed-script named entities. We evaluate five multilingual ASR models on two benchmarks using an 11-category error taxonomy. Script confusion drives the majority of Whisper's failures on three of four languages. SeamlessM4T achieves the best results (16.3% WER on Urdu, 22.1% on Punjabi), though still far from production-grade accuracy. Cross-language contamination affects up to 35% of output characters in lower-resource languages; rare words reach only 20–62% accuracy; and short utterances are harder than long ones, reversing the high-resource pattern. Standard WER conflates recognition errors with orthographic and Unicode variation, overstating rates by up to 3.1 points. Transliteration post-processing recovers up to 25 points of WER without retraining. Our findings provide actionable guidance for ASR in this language family.

Index Terms: speech recognition, low-resource languages, error analysis, South Asian languages, multilingual ASR

1. Introduction

Recent multilingual ASR models claim coverage of 100+ languages [4, 14, 15], yet the quality of that coverage varies enormously. For well-resourced languages, WER has dropped below 5%; for many others, models are deployed with little understanding of how or why they fail. This gap matters most where speech technology could have the greatest impact: healthcare delivery, government services, and digital access in regions where literacy rates are low and voice is the primary interface.

Urdu, Punjabi, Pashto, and Sindhi are spoken by over 350 million people across Pakistan, India, and Afghanistan. They share geographic and cultural proximity but differ in script, resource availability, and morphological structure. Critically, they present a combination of challenges rarely encountered together in well-studied language families. Urdu, Pashto, and Sindhi all use Arabic script variants with overlapping but distinct character inventories, while Punjabi uses Gurmukhi but shares vocabulary with Hindi (Devanagari), creating opportunities for cross-script confusion at the model level. Unicode encodes visually identical characters with different codepoints, inflating error rates without reflecting recognition failure. And resource levels vary dramatically within the family: Urdu has hundreds of hours of training data while Sindhi only has tens of hours, changing the *nature* of errors, not just their frequency.

Prior evaluations of ASR for these languages report aggregate WER on one or two benchmarks [3, 13, 17], collapsing all these phenomena into a single number. This obscures whether poor performance reflects acoustic recognition difficulty, script-level decoding failures, or evaluation artifacts. We address this gap with three contributions:

1. A comprehensive benchmark evaluating five models spanning three architecture families (Whisper [15], MMS [14], SeamlessM4T [4]) across 4 languages on 2 datasets (40 configurations).
2. An 11-category error taxonomy that decomposes aggregate WER into interpretable, actionable dimensions.
3. Concrete findings: script confusion drives Whisper's WER >100% on three languages; simple post-processing recovers up to 25 pp WER; Unicode normalization reduces Pashto WER by 3.1 pp; named entities lag non-entity words by 5–20 pp; and resource level changes the nature of failures, not just their frequency.

Table 1: *Languages, datasets, and sample utterances.*

Language	Script	Resource	CV	FLEURS
Urdu	Arabic	Mid	5,082	299
Punjabi	Gurmukhi	Low	587	574
Pashto	Arabic	Low	3,610	512
Sindhi	Arabic	Very Low	40	980

Lang	Sample utterance
Urdu	ماحولیاتی آلودگی اور پوری دنیا کے بڑھتے ہوئے درجۂ حرارت
Punjabi	ਗੁਆਂਢੀਆਂ ਨੂੰ ਕਦੇ ਵੀ ਮਾੜਾ ਲਫ਼ਜ਼ ਨਾ ਬੋਲੋ
Pashto	ایا تاسو د پیدل سفر لپاره کوم وړاندیز لرئ؟
Sindhi	دلچسپ جملن کي ترتيب ڏيڻ سندن پسنديده تفريح آهي

2. Experimental Setup

2.1. Languages and Datasets

Table 1 summarizes the four target languages and evaluation datasets. We use Common Voice 22.0 [2] and FLEURS [6] to evaluate cross-benchmark consistency. Test set sizes range from 40 utterances (Sindhi CV) to 5,082 (Urdu CV), reflecting the extreme resource disparity within this language family. The dialectal composition of these benchmarks remains largely undocumented in the literature: Common Voice relies on optional, self-reported accent metadata with sparse coverage, and FLEURS provides no dialect scope information for any language subset [6, 11].

Table 2: *WER (%) across all configurations. Best in **bold**. WER >100% indicates wrong-script output.*

			SeamlessM4T		Whisper	
		MMS	**M**	**L**	**med**	**lg-v3**
Urdu	CV	31.8	22.8	**17.1**	35.4	21.2
	FL	29.7	18.6	**16.3**	28.3	21.6
Punjabi	CV	33.8	39.6	**28.6**	104.7	85.3
	FL	30.2	29.1	**22.1**	101.2	87.8
Pashto	CV	63.1	61.4	**42.0**	165.3	90.4
	FL	43.9	57.6	**43.8**	103.5	89.2
Sindhi	CV	**37.5**	59.5	55.9	144.4	103.0
	FL	**23.5**	60.2	55.5	109.9	100.7

2.2. Models

We evaluate five models spanning three architecture families: Whisper-medium (769M) and Whisper-large-v3 (1.5B) [15], autoregressive encoder-decoders; MMS-1b (1B) [14], a CTC model [9]; and SeamlessM4T-Medium (1.2B) and SeamlessM4T-Large (2.3B) [4], encoder-decoders with language adapters. These architectures differ in their capacity for hallucination: CTC models cannot generate tokens beyond the input length, autoregressive models can hallucinate freely, and language adapters provide script-level guidance.

2.3. Evaluation Framework

We compute WER and CER using jiwer [10] as the baseline metric. We then apply an 11-category error taxonomy post-hoc, covering script confusion, orthographic variants, character-level confusions, named entity accuracy, OOV/rare word effects, utterance length degradation, cross-language interference, numeral handling, transliteration patterns, edit operation profiles, and repetition/looping behavior. Named entities are annotated using an LLM-based pipeline (Claude Haiku 4.5 (`claude-haiku-4-5-20251001`)) with four entity types (PER, LOC, ORG, MISC); entity accuracy is defined in Section 3.4. Word frequency is computed from Common Voice training-set transcripts.

3. Multi-Dimensional Error Analysis

Table 2 presents WER across all 40 configurations. SeamlessM4T-Large achieves the best WER on 6 of 8 language–dataset pairs (MMS-1b wins both Sindhi pairs). Table 3 provides a roadmap of the 11 error dimensions we decompose below.

3.1. Script Confusion

The most dramatic failures in our evaluation stem from *script confusion*: models outputting text in the wrong writing system entirely.

Wrong-script output. Whisper-medium produces problematic output for 100% of Sindhi utterances (63% Devanagari, 37% garbage) and for 95% of Punjabi utterances on Common Voice, despite these languages using Arabic and Gurmukhi scripts respectively. For Pashto, 80% of Whisper-medium's Common Voice outputs are problematic: 22% wrong-script and 58% garbage (Khmer script, invisible characters). Whisper-large-v3 shows similar patterns at reduced rates. In contrast, MMS-1b and SeamlessM4T produce near-zero script confusion across all languages.

Table 3: *Error taxonomy summary with key findings across 11 dimensions.*

Dimension	**Key Finding**
Script confusion	Whisper: 80–100% wrong script on 3/4 languages; transliteration recovers 7–25 pp WER
Edit operations	MMS: 77–90% subs; Whisper: elevated insertions
Repetitions	CTC immune; autoregressive loops WER >200%
Numerals	Each architecture has a different failure mode
Orthographic var.	Multiple valid spellings penalized as errors
Char. confusions	Pashto Yeh: 11,547 Unicode artifact instances
Named entities	5–20 pp harder than non-entities; 0–80% by model
Latin tokens	27–62% deleted; transliterations penalized
OOV/rare words	30–45 pp gap rare vs. common
Length	Short > long WER (reversed pattern)
Cross-language	Sindhi: 12–35% contamination from Urdu

WER inflation from script confusion. Removing script-confused outputs reduces Whisper-medium's Pashto CV WER from 165.3% to approximately 97%, representing a 68 percentage point inflation from script confusion alone. WER exceeding 100% is a reliable signal that a model is generating in the wrong language, not merely producing poor transcriptions.

Post-processing transliteration. Since Whisper's wrong-script outputs are phonetically plausible, we test whether simple script conversion can recover usable transcriptions. We apply rule-based transliteration to Whisper predictions using aksharamukha [16] (Devanagari→Gurmukhi for Punjabi) and indo-arabic-transliteration [8] (Devanagari→Sindhi Arabic). On Common Voice, this reduces Whisper-medium WER by 25.0 pp on Punjabi (104.7→79.7%) and CER by 62.5 pp, confirming that the underlying phonetic recognition is largely intact. Sindhi WER drops by 7.6 pp with CER falling 33.4 pp. Whisper-large-v3 shows smaller but consistent gains (-13.3 pp WER, -49.3 pp CER on Sindhi). These improvements require no retraining, only string post-processing, and establish a lower bound on recoverable accuracy lost to script confusion.

3.2. Architecture-Dependent Failure Modes

Beyond script confusion, model architecture shapes the qualitative nature of errors.

Edit operation profiles. Decomposing errors into substitutions, deletions, and insertions reveals qualitatively different failure modes (Figure 1). MMS-1b shows a substitution-dominated profile (77–90% substitutions, 8–19% deletions, 2–8% insertions), consistent with its CTC architecture that cannot insert tokens beyond the input. Whisper-medium shows markedly elevated insertions (up to 23% of errors), reflecting hallucinated content from its autoregressive decoder. SeamlessM4T falls between these extremes.

Language-specific error profiles. Breaking down error composition by language reveals that the averaged view masks substantial per-language variation. Whisper-medium's insertion rate on Pashto reaches 27%, nearly double its average, while on Urdu it drops to 13%. MMS-1b's deletion rate on Pashto (19%) is nearly twice its Urdu rate (10%), suggesting that lower-resource languages trigger qualitatively different failure modes

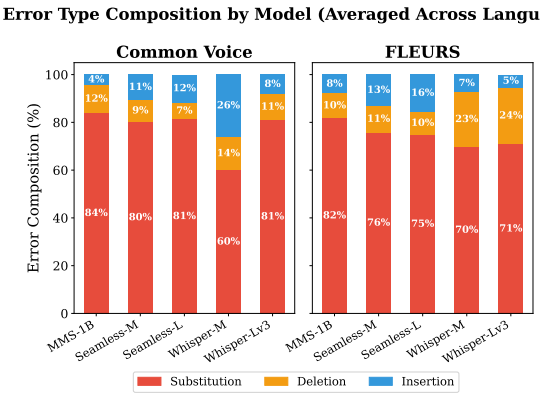

Figure 1: *Error type composition by model, averaged across languages. MMS is substitution-dominated (CTC); Whisper shows elevated insertions.*

even within the same architecture.

Repetition and looping. CTC-based MMS is architecturally immune to repetition loops (0 instances). Autoregressive models (Whisper, SeamlessM4T) occasionally enter decoding loops, producing catastrophic outputs with WER exceeding 200%.

Numeral handling. On FLEURS, SeamlessM4T spells out 87–89% of numerals in words; MMS attempts digit form (23% correct, 44% incorrect); Whisper drops 21–27% of numerals entirely. No model reliably handles numerals, and each architecture exhibits a distinct failure mode.

3.3. Orthographic Variation and Unicode Ambiguity

A substantial fraction of measured WER reflects not recognition failure but ambiguity inherent in these writing systems.

Unicode normalization artifacts. The top character "confusion" in Pashto is Arabic Yeh ي (U+064A) versus Farsi Yeh ی (U+06CC), totaling 11,547 instances across all Pashto configurations. These characters are visually identical but occupy different Unicode codepoints, inflating WER without reflecting any acoustic recognition error [19]. Similar issues affect Hamza variants, Arabic Kaf ک versus Keheh ڪ, and Heh ہ versus Heh Goal ۂ.

Orthographic variants. Punjabi exhibits legitimate spelling alternations penalized by WER: gemination markers (e.g., *vich* vs. *vichchh*, with and without gemination diacritic), and aspirate variants. In Pashto and Urdu, optional characters produce valid alternatives (e.g., multiple spellings of "Afghanistan") that WER treats as errors.

Cross-script character leakage. Punjabi shows Gurmukhi-to-Devanagari character-level confusion in hundreds of instances (e.g., ਰ → र, ◌ਾ → ◌ा), where models substitute visually similar characters across related scripts.

These findings support reporting both raw and Unicode-normalized WER: normalization better reflects end-user experience, while raw WER surfaces preprocessing issues for indexing.

Quantifying the fix: WER after Unicode normalization. To move beyond diagnosis, we implement a Perso-Arabic character normalization step that unifies visually identical codepoint variants (Yeh, Kaf, Heh, Hamza, and Alef variants) before WER computation. Table 4 reports the impact on Common Voice WER. Pashto benefits most (−3.1 pp average), consistent with the 11,547 Yeh-variant instances documented above. Punjabi sees −1.6 pp and Urdu −0.4 pp. The effect is largest for models that otherwise produce correct-script output (MMS, SeamlessM4T), where Unicode artifacts constitute a larger share of measured errors. For Whisper on Sindhi and Punjabi, normalization has negligible effect because the dominant errors are script confusion, not character variants. We recommend this normalization as a standard preprocessing step for Perso-Arabic ASR evaluation.

Table 4: *Average WER and CER reduction (pp) on Common Voice after Perso-Arabic Unicode normalization (unifying Yeh, Kaf, Heh, Hamza variants).*

Language	Avg ΔWER	Avg ΔCER
Pashto	−3.1	−1.9
Punjabi	−1.6	−1.1
Sindhi	−0.8	−0.3
Urdu	−0.4	−0.2

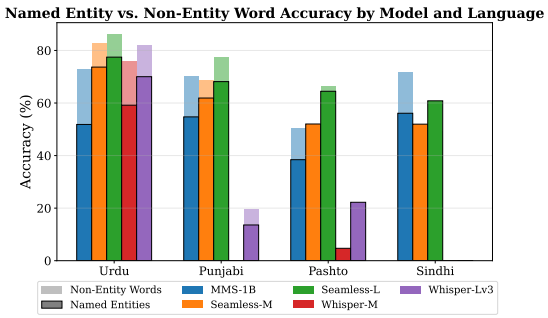

Figure 2: *Named entity vs. non-entity accuracy by model and language. Entities are consistently harder for lower-resource languages.*

3.4. Named Entities and Latin Token Handling

Named entities (proper nouns, locations, organizations) pose distinct challenges in multilingual and multi-script contexts.

Entity accuracy varies dramatically by architecture. Entity accuracy is the exact-match rate of LLM-annotated reference entity tokens under jiwer alignment. SeamlessM4T-Large achieves 65–80% entity accuracy, MMS-1b achieves 30–70%, and Whisper-large-v3 achieves 0–72% (near-zero except Urdu). Entities are consistently harder than non-entity words, with a 5–20 percentage point accuracy gap in most configurations (Figure 2). Whisper's entity failures on Sindhi and Punjabi trace entirely to script confusion (Section 3.1): when the output is in the wrong script, entity accuracy drops to near zero.

Latin token handling. Models overwhelmingly transliterate or drop Latin loanwords present in references: 27–62% are deleted, 9–54% transliterated to native script, and only 1–11% preserved as-is. SeamlessM4T prefers transliteration; MMS prefers deletion. Critically, WER penalizes correct transliterations (e.g., "scroll" → native-script equivalent), revealing a fundamental limitation of character-matching evaluation for code-mixed content.

3.5. Resource Level Effects

The four languages span a resource continuum from mid-resource (Urdu) to very low-resource (Sindhi). We find that resource level changes the *nature* of failures, not just their frequency.

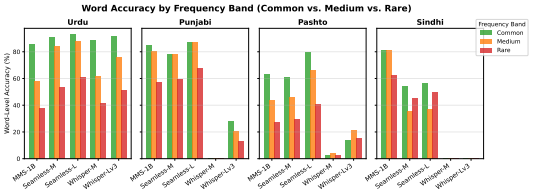

Figure 3: *Word accuracy by frequency bin (rare/medium/common). All models show a frequency–accuracy gradient, steepest for better-resourced languages.*

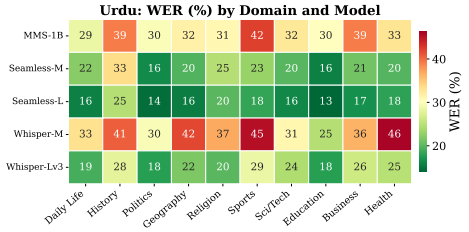

Figure 4: *WER (%) by domain and model for Urdu. Domain difficulty varies by 8–14 pp; the hardest domains differ by model.*

Word frequency gradient. Using word frequencies computed from Common Voice training transcripts (Urdu: 7,326 sentences; Pashto: 4,611; Punjabi: 800; Sindhi: 271), we observe a clear frequency–accuracy gradient (Figure 3). Rare words ($\leq$2 occurrences in training) are recognized at 20–62% accuracy versus 49–94% for common words ($>$10 occurrences). The gradient is steepest for Urdu (30–45 pp gap) and Pashto (30–40 pp gap), and persists even for the lowest-resource Sindhi on FLEURS.

Reversed length effect. Short utterances (1–5 words) yield approximately 76% mean WER versus 46% for long utterances (21+ words), reversing the typical high-resource pattern. In per-length bins, a single substitution in a one-word utterance yields 100% WER, amplifying short-utterance error rates.

Cross-language contamination. Sindhi outputs contain 12–35% characters from other languages (primarily Urdu), while Urdu shows near-zero contamination. Contamination rate correlates inversely with resource level: the less training data available, the more models "borrow" from better-resourced related languages (e.g., 1,600+ hours of Hindi [5] vs. Sindhi's 271 training sentences).

Domain-level variation. We classify FLEURS and Common Voice utterances into 10 topic domains using an LLM classifier and analyze domain-level WER for Urdu (Figure 4), the only language where all models produce correct-script output, isolating genuine content difficulty from script confusion. WER varies by 8–14 pp across domains: *education* and *politics* are easiest (20–22% average), while *history/culture* and *sports* are hardest (31–33%), likely reflecting specialized vocabulary absent from training data. Notably, the hardest domain differs by model, likely reflecting differences in pretraining data composition.

4. Related Work

Arif et al. [3] benchmarked Whisper, MMS, and SeamlessM4T on Urdu and identified text normalization as a key confound. Our Unicode analysis (Section 3.3) confirms and quantifies this across four languages. Sehar et al. [17] evaluated Whisper on Pashto, Punjabi, and Urdu with n-shot prompting but reported only aggregate WER; our error taxonomy reveals that Whisper's poor scores on these languages are predominantly script confusion rather than acoustic failure, a distinction invisible to WER alone. Naseem et al. [13] benchmarked MMS for Punjabi and Urdu TTS, highlighting the limited speech technology infrastructure for these languages. Our work extends both studies by covering more models, more languages, and providing the first systematic decomposition of *why* models fail, not just *how much*. Prior ASR error analyses have categorized failures for single high-resource languages [20], and recent work has shown that WER itself is an unreliable metric across writing systems [18]. Our 11-category taxonomy retains standard dimensions (edit operations, OOV effects) while adding categories specific to multi-script low-resource settings: script confusion, Unicode normalization artifacts, cross-language contamination, and orthographic variation. Similar benchmarking for African languages reveals parallel failure modes: Nahabwe et al. [12] evaluate four ASR systems across 13 African languages and find no single model dominates across resource levels, while Diallo et al. [7] report Whisper variants exceeding 100% WER on Bambara, the same catastrophic pattern we observe on Whisper.

5. Discussion

Our 11-dimensional analysis (Table 3) reveals that aggregate WER systematically misattributes the nature of failures for South Asian ASR. The dominant challenges are structural: script confusion, Unicode inconsistencies, orthographic instability, and cross-language interference.

For model developers. Script confusion is Whisper's dominant failure mode on three of four languages, but it is entangled with deeper model uncertainty for lowest-resource languages, so decoding-time fixes alone are unlikely to suffice. Post-processing transliteration recovers up to 25 pp WER without retraining. Language adapter architectures (SeamlessM4T) are the most robust for multi-script families.

For evaluation. Report both raw and Unicode-normalized WER for Perso-Arabic scripts, where Yeh/Kaf/Heh variants inflate error rates by up to 3.1 pp (Table 4). Named entity accuracy and per-length-bin WER provide more informative diagnostics than aggregate WER alone.

For deployment. The best FLEURS results (Urdu 16.3%, Punjabi 22.1%, Sindhi 23.5%, Pashto 43.8%) remain well above the sub-10% WER target for production-grade ASR in domains such as healthcare, and all are on read-aloud speech. Real deployments with spontaneous, code-mixed speech would yield substantially higher error rates.

For the community. These challenges (script proximity, orthographic instability, resource disparity) likely generalize to other Perso-Arabic and Indic script families. Crowdsourcing initiatives such as Karya [1] could help close data gaps for Sindhi and Pashto.

Limitations. Sindhi Common Voice contains only 40 test utterances, limiting statistical reliability. We evaluate only zero-shot inference; fine-tuning would likely reduce many failure modes documented here.

Conclusion. Our 11-dimensional analysis across 40 configurations reveals that script confusion dominates Whisper's errors; that orthographic variants and Unicode artifacts are major confounds for WER; and that resource level changes the nature of failures, not just their frequency. We release our evaluation framework and error taxonomy to support targeted improvements for ASR in this underserved language family.

6. Use of Generative AI Disclosure

Generative AI tools were used only to assist in polishing the manuscript text. Experimental design, analysis, and scientific conclusions are the sole work of the authors.

7. References

[1] Basil Abraham, Danish Goel, Divya Siddarth, Kalika Bali, Manu Chopra, Monojit Choudhury, Pratik Joshi, Preethi Jyoti, Sunayana Sitaram, and Vivek Seshadri. Crowdsourcing speech data for low-resource languages from low-income workers. In *Proceedings of the 12th Language Resources and Evaluation Conference (LREC)*, pages 2819–2826, Marseille, France, 2020. European Language Resources Association. URL https://aclanthology.org/2020.lrec-1.343/.

[2] Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis Tyers, and Gregor Weber. Common voice: A massively-multilingual speech corpus. In *Proceedings of the 12th Language Resources and Evaluation Conference (LREC)*, pages 4218–4222, 2020.

[3] Samee Arif, Aamina Jamal Khan, Mustafa Abbas, Agha Ali Raza, and Awais Athar. WER we stand: Benchmarking Urdu ASR models. In *Proceedings of the 31st International Conference on Computational Linguistics (COLING)*, 2025.

[4] Loïc Barrault, Yu-An Chung, Mariano Coria Meglioli, David Dale, Ning Dong, Paul-Ambroise Duquenne, Hady Elsahar, Hongyu Gong, Kevin Heffernan, John Hoffman, et al. Seamlessm4t—massively multilingual & multimodal machine translation. *arXiv preprint arXiv:2308.11596*, 2023.

[5] Kaushal Santosh Bhogale, Abhigyan Raman, Tahir Javed, Sumanth Doddapaneni, Anoop Kunchukuttan, Pratyush Kumar, and Mitesh M. Khapra. Effectiveness of mining audio and text pairs from public data for improving ASR systems for low-resource languages. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023.

[6] Alexis Conneau, Min Ma, Simran Khanuja, Yu Zhang, Vera Axelrod, Siddharth Dalmia, Jason Riesa, Clara Rivera, and Ankur Bapna. FLEURS: Few-shot learning evaluation of universal representations of speech. In *Proceedings of the IEEE Spoken Language Technology Workshop (SLT)*, pages 798–805, 2023.

[7] Seydou Diallo, Yacouba Diarra, Mamadou K. Keita, Panga Azazia Kamaté, Adam Bouno Kampo, and Aboubacar Ouattara. Where are we at with automatic speech recognition for the Bambara language? *arXiv preprint arXiv:2602.09785*, 2026. URL https://arxiv.org/abs/2602.09785.

[8] GokulNC. Indo-arabic transliteration: Script conversion for indo-pakistani languages, 2023. URL https://github.com/GokulNC/Indic-PersoArabic-Script-Converter.

[9] Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd International Conference on Machine Learning (ICML)*, pages 369–376, 2006.

[10] Jitsi. jiwer: Evaluate your speech-to-text system with similarity measures such as word error rate (WER), 2024. URL https://github.com/jitsi/jiwer.

[11] Mingfei Lau, Qian Chen, Yeming Fang, Tingting Xu, Tongzhou Chen, and Pavel Golik. Data quality issues in multilingual speech datasets: The need for sociolinguistic awareness and proactive language planning. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7466–7492, Vienna, Austria, 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.acl-long.370. URL https://aclanthology.org/2025.acl-long.370/.

[12] Alvin Nahabwe, Sulaiman Kagumire, Denis Musinguzi, Bruno Beijuka, Jonah Mubuuke Kyagaba, Peter Nabende, Andrew Katumba, and Joyce Nakatumba-Nabende. Benchmarking automatic speech recognition models for african languages. *arXiv preprint arXiv:2512.10968*, 2025. URL https://arxiv.org/abs/2512.10968.

[13] Fatima Naseem, Maham Sajid, Farah Adeeba, Sahar Rauf, Asad Mustafa, and Sarmad Hussain. Developing high-quality TTS for Punjabi and Urdu: Benchmarking against MMS models. In *Proceedings of Interspeech*, pages 1628–1632, 2025. doi: 10.21437/Interspeech.2025-469.

[14] 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. Scaling speech technology to 1,000+ languages. *Journal of Machine Learning Research*, 25:1–52, 2024.

[15] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, pages 28492–28518, 2023.

[16] Vinodh Rajan. Aksharamukha: Script converter, 2020. URL https://github.com/virtualvinodh/aksharamukha.

[17] Najm Ul Sehar, Ayesha Khalid, Farah Adeeba, and Sarmad Hussain. Benchmarking Whisper for low-resource speech recognition: An n-shot evaluation on Pashto, Punjabi, and Urdu. In *Proceedings of the First Workshop on Challenges in Processing South Asian Languages (CHiPSAL 2025), COLING*, pages 202–207, 2025.

[18] D K Thennal, Jesin James, Deepa Padmini Gopinath, and Muhammed Ashraf K. Advocating character error rate for multilingual ASR evaluation. In *Findings of the Association for Computational Linguistics: NAACL 2025*, 2025. URL https://aclanthology.org/2025.findings-naacl.277/.

[19] Jan Trmal, Gaurav Kumar, Vimal Manohar, Sanjeev Khudanpur, Matt Post, and Paul McNamee. Using of heterogeneous corpora for training of an ASR system, 2017. URL https://arxiv.org/abs/1706.00321.

[20] Johannes Wirth and Rene Peinl. ASR in German: A detailed error analysis. *arXiv preprint arXiv:2204.05617*, 2022. URL https://arxiv.org/abs/2204.05617.