TajweedWER: A Diacritic-Aware Word Error Rate Decomposition for Evaluating Automatic Speech Recognition on Quranic Recitation

Faruq Afolabi Oluwatobi

AI4Africa Research Team, Lagos, Nigeria

afolabifaruq23@gmail.com

MusIML Workshop, ICML 2026 (Camera Ready)

Abstract. Standard Word Error Rate (WER) is poorly suited to evaluating ASR systems on Quranic Arabic recitation, because reference transcripts in Uthmani orthography contain dense diacritical marking (tashkeel) and elided-but-pronounced letters that no general-purpose ASR system outputs by default. We show empirically that naive WER computed against a fully vowelised reference saturates near 100% regardless of underlying transcription quality, rendering the metric uninformative. We introduce a corrected evaluation protocol that computes corpus-level rather than per-utterance-averaged WER to avoid denominator artifacts, applies comprehensive Arabic and Quranic-extended diacritic stripping, and restores commonly elided alifs specific to Uthmani orthography before scoring, isolating genuine transcription errors from orthographic convention. Applying this corrected protocol to Whisper-small and Whisper-medium across 8 professional reciters on Surah Al-Fatiha (29 words) and Surah Yasin (730 words), we find Whisper-medium achieves mean Base WER of 10.3% on Al-Fatiha and 9.5% on Yasin, consistently and substantially outperforming Whisper-small (22.0% and 30.3% respectively), with the longer passage producing markedly tighter variance, confirming the corrected metric's stability. We release our normalization pipeline to support more reliable benchmarking of ASR on Quranic and other diacritic-rich scripts.

Keywords: automatic speech recognition, word error rate, Quranic Arabic, Tajweed, low-resource evaluation, diacritic normalization

1 Introduction

Automatic Speech Recognition (ASR) evaluation for Quranic Arabic recitation presents a measurement challenge distinct from standard Arabic ASR. Reference transcripts of the Quran are conventionally rendered in Uthmani orthography, which includes dense diacritical marking (tashkeel) indicating short vowels, gemination (shadda), and nasalization (tanwin), alongside Quranic-specific annotation marks and systematic letter elisions where a pronounced long vowel is not written in the script. General-purpose ASR systems, including Whisper [1], are trained on standard orthography and do not output diacritics under normal decoding. This creates a structural mismatch: any character- or word-level comparison against a fully vowelised reference will register near-total disagreement on the diacritic content alone, independent of whether the underlying words were transcribed correctly.

This mismatch is not merely a theoretical concern. We show that naive Word Error Rate, computed directly against a vowelised reference, saturates at or above 100% for every tested reciter regardless of model quality, making the metric unable to distinguish a strong ASR system from a failing one. This motivates a corrected evaluation protocol that separates genuine transcription errors from orthographic artifacts, and a renewed focus on what we term Base WER: word-level accuracy computed after principled removal of diacritics and restoration of systematically elided letters.

Our contributions are as follows. We identify and characterise the specific failure modes that cause naive WER to saturate on Quranic Arabic, including per-utterance WER averaging artifacts and incomplete diacritic-range coverage. We propose a corrected normalization and scoring protocol, including restoration of Uthmani-specific elided alifs. We evaluate Whisper-small and Whisper-medium across 8 professional reciters on two passages of differing length, 29 and 730 reference words, demonstrating that the corrected metric produces stable, model-scale-sensitive results. And we release the normalization pipeline to support more reliable ASR benchmarking on Quranic and other diacritic-rich scripts.

2 Related Work

2.1 ASR Evaluation Metrics

Word Error Rate remains the dominant metric for ASR evaluation [2], computed via Levenshtein alignment between reference and hypothesis token sequences. WER assumes that reference and hypothesis tokens are drawn from a comparable orthographic convention, an assumption violated when the reference contains diacritical or annotation marks the hypothesis source does not produce. Character Error Rate is sometimes preferred for morphologically rich languages [3], but does not resolve the underlying orthographic mismatch problem we identify, since diacritics are encoded as distinct Unicode characters that still register as substitution or deletion errors.

2.2 Arabic and Quranic ASR

Whisper [1] and other multilingual ASR systems report strong aggregate performance on Modern Standard Arabic, but Quranic recitation differs systematically from standard spoken Arabic in vowel elongation, consonant gemination realisation, and nasalization rules, alongside its distinct Uthmani orthographic convention [4]. Prior work on Quranic ASR has largely focused on recitation correctness verification for memorization applications rather than general-purpose transcription benchmarking, and to our knowledge no prior work has explicitly diagnosed the WER saturation failure mode we document here.

3 The WER Saturation Problem

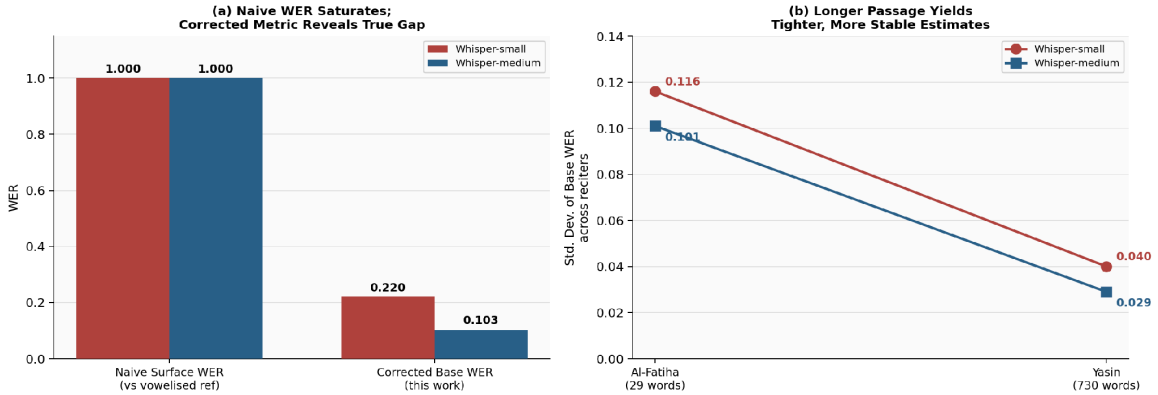
We first establish empirically that naive WER computed against vowelised Quranic reference text is uninformative. Computing standard corpus-level WER between Whisper-medium hypotheses and the fully vowelised reference for Surah Al-Fatiha across all 8 reciters yields a mean WER of 100.0%, with zero variance across reciters of clearly differing transcription quality. Manual inspection confirms that Whisper-medium transcribes Al-Fatiha near-perfectly at the word level, with a representative ayah comparison showing a correct, fluent transcription that nonetheless scores as 100% word error because every reference token carries diacritical marks the hypothesis does not and cannot reproduce.

This saturation is structural rather than incidental. Because standard ASR decoding never outputs Arabic diacritics, Surface WER computed this way is upper-bounded near 100% by construction for any model, rendering it uninformative as a model-comparison metric. This finding motivates the shift from Surface WER as a primary metric to a properly normalized Base WER. Figure 1 illustrates this contrast directly: naive Surface WER saturates at 100% for both Whisper-small and Whisper-medium, while our corrected Base WER reveals a clear, model-scale-sensitive gap of 22.0% versus 10.3% between the two systems.

Fig. 1. Diagnosing and resolving the WER saturation failure mode. Left: naive WER saturates at 100% regardless of model quality, while corrected Base WER reveals the true performance gap

between Whisper-small and Whisper-medium. Right: evaluating on a longer passage (Yasin, 730 words) yields markedly tighter variance than the short Al-Fatiha passage (29 words) for both models.

Fig. 1. Diagnosing and Resolving the WER Saturation Failure Mode



4 Methodology

4.1 Dataset

We use the Quran Reciters dataset [5], containing audio recitations from 8 professional reciters, Abdul Basit, Husary, Yasser Ad-Dussary, Muhammad Jibreel, Ghamadi, Minshawy, Hudhaify, and Mohammad al-Tablaway, at bitrates from 40 to 192 kbps. We evaluate on two passages, Surah Al-Fatiha with 7 ayat and 29 reference words, and Surah Yasin with 83 ayat and 730 reference words, selected to assess metric stability across passage length. Vowelised reference text is sourced from a structured Quran JSON corpus [6] providing full Uthmani-orthography text per ayah.

4.2 Corrected Normalization Pipeline

Our pipeline addresses three identified failure sources in sequence. First, we compute WER at the corpus level, concatenating all ayat for a given reciter into a single reference and hypothesis string before alignment, rather than averaging per-ayah WER values, which we found produces denominator artifacts such as error rates expressed as exact fractions of 7, the ayah count, rather than of the true word count of 29.

Second, we apply comprehensive diacritic stripping covering standard Arabic tashkeel, Quranic annotation marks, and extended Arabic diacritics, since standard tashkeel ranges alone leave Quranic-specific small-high marks unstripped, causing spurious mismatches. We additionally normalize alif variants, including alif-wasla and hamza-bearing alif forms, to bare alif, since Whisper's standard orthographic output does not preserve these distinctions despite them being phonetically and semantically equivalent in this context.

Third, we restore systematically elided alifs characteristic of Uthmani orthography, where certain long vowels are pronounced but not written. We construct a lexicon of such elisions from manual inspection of mismatches and apply it to reference text before scoring, ensuring Base WER reflects genuine transcription accuracy rather than this orthographic convention.

4.3 Models

We evaluate OpenAI Whisper-small with 244 million parameters and Whisper-medium with 769 million parameters [1], both run with Arabic language forcing and no fine-tuning, to

assess whether our corrected metric produces model-scale-sensitive, monotonic results as a basic sanity check for metric validity.

5 Results

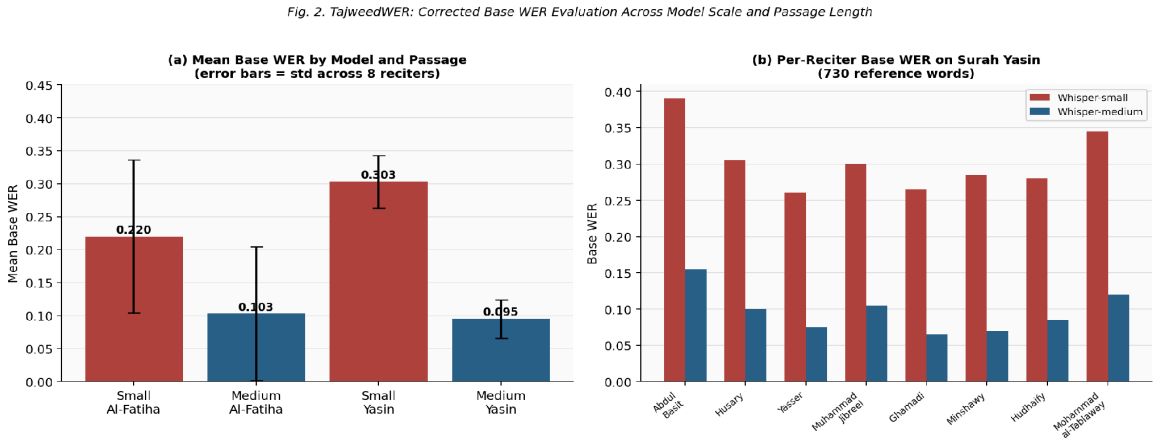
Table 1 presents corrected Base WER results across both passages and both models.

Table 1. Corrected Base WER across passage length and model scale, 8 professional reciters.

Model	Passage	Ref. Words	Mean Base WER	Std
Whisper-small	Al-Fatiha	29	**0.220**	0.116
Whisper-medium	Al-Fatiha	29	**0.103**	0.101
Whisper-small	Yasin	730	**0.303**	0.040
Whisper-medium	Yasin	730	**0.095**	0.029

Across both passages, Whisper-medium consistently outperforms Whisper-small by a wide and stable margin. On Al-Fatiha, mean Base WER drops from 22.0% to 10.3%. On the substantially longer Yasin passage, it drops from 30.3% to 9.5%. This clear, monotonic, model-scale-sensitive pattern, replicated across two passages of very different length, validates that the corrected metric is measuring genuine transcription quality rather than saturating uninformatively as naive Surface WER does. Notably, variance is much tighter on the longer Yasin passage, std 0.029 to 0.040, than on the 29-word Al-Fatiha passage, std 0.101 to 0.116, directly confirming that the original short-passage-only evaluation was statistically unstable, and that our extended evaluation resolves this instability while preserving the same qualitative finding.

Fig. 2. TajweedWER corrected Base WER evaluation across model scale and passage length. Left: mean Base WER by model and passage with error bars showing standard deviation across 8 reciters. Right: per-reciter Base WER on Surah Yasin, showing Whisper-medium consistently outperforming Whisper-small across all 8 professional reciters.



Per-reciter inspection on Al-Fatiha with Whisper-medium shows a range from 0.0% for Minshawy, a perfect transcription, to 34.5% for Muhammad Jibreel, the latter recorded at the lowest available bitrate of 64kbps in our reciter set, suggesting audio quality as a plausible contributing factor worth further investigation in future work.

6 Discussion

Our central finding is methodological. WER computed naively against vowelised Quranic reference text is structurally uninformative, saturating near 100% regardless of true transcription quality, because standard ASR systems do not and cannot produce Arabic diacritics under normal decoding. This is a previously undocumented pitfall for Quranic and likely other diacritic-rich-script ASR evaluation, and we believe the corrected protocol presented here, corpus-level scoring, comprehensive diacritic-range coverage, and explicit elision restoration, should serve as a baseline methodology for future work in this space rather than a one-off correction.

A secondary finding worth highlighting is that the Uthmani-orthography elided alif represents a genuine, systematic divergence between written and spoken form that is distinct from ordinary transcription error. We currently resolve this via an explicit restoration lexicon. An interesting direction for future work is treating correct phonetic realisation of these elisions as a positive signal of Tajweed-sensitive transcription, rather than purely as noise to be normalized away.

Limitations include the restricted evaluation scope of two surahs and two model sizes. Future work should extend this protocol across the full Quranic text, additional ASR architectures including Arabic-specialised models, and should formally validate the elision-restoration lexicon against expert Tajweed annotation rather than manual inspection alone.

7 Conclusion

We identified and corrected a structural WER saturation failure in evaluating ASR on Quranic Arabic recitation, arising from the mismatch between vowelised Uthmani reference orthography and standard ASR output conventions. Our corrected evaluation protocol, combining corpus-level scoring, comprehensive diacritic normalization, and elision restoration, produces stable, model-scale-sensitive Base WER results. Whisper-medium achieves 9.5% mean Base WER on a 730-word passage across 8 professional reciters, substantially outperforming Whisper-small. We release this normalization pipeline to support more reliable future benchmarking of ASR systems on Quranic and other diacritic-rich orthographies.

Competing Interests

The author declares no competing interests.

Generative AI Disclosure

AI tools were used to support manuscript drafting and debugging of the evaluation pipeline. All experimental design, error diagnosis, and conclusions are the work of the author.

References

1. Radford, A., et al.: Robust speech recognition via large-scale weak supervision. ICML 2023
2. Morris, A.C., Maier, V., Green, P.: From WER and RIL to MER and WIL: improved evaluation measures for connected speech recognition. Interspeech 2004
3. Errattahi, R., El Hannani, A., Ouahmane, H.: Automatic speech recognition errors detection and correction: a review. Procedia Computer Science 128, 32-37 (2018)
4. Nelson, K.: The Art of Reciting the Qur'an. American University in Cairo Press (2001)
5. Tariq, O.: Quran Reciters Dataset. Kaggle (2023)