# Qur'anic Phonemizer: Bringing Tajweed-Aware Phonemes to Qur'anic Machine Learning

**Ahmed Ibrahim, Mostafa Shahin, Beena Ahmed**
The University of New South Wales
ahmed.ibrahim8165@gmail.com, m.shahin@unsw.edu.au, beena.ahmed@unsw.edu.au

## Abstract

Qur'anic recitation follows explicit Tajweed rules that standard Arabic grapheme-to-phoneme tools do not capture, limiting phoneme-level research for the Qur'an. We introduce a modular, computationally efficient Python API for the Hafs 'an Asim recitation style that converts Qur'anic text into a configurable 71-symbol phoneme inventory, comprehensively encoding Tajweed rules such as Idgham, Iqlab, Ikhfaa, Qalqala, Tafkheem, Waqf, etc. We anticipate that this tool will have various use cases in speech recognition, mispronunciation detection, text-to-speech, linguistic analysis and pedagogical applications to name a few. Current limitations include support for Hafs only—extensions to other recitation styles are discussed. The code (https://github.com/Hetchy/Quranic-Phonemizer) and user interface (https://quranicphonemizer.com) are released as open source.

## 1 Introduction

Phonemes are the smallest units of sound. A phonemizer converts written text (graphemes) into phoneme sequences, which is commonly used in Automatic Speech Recognition (ASR) and Mispronunciation Detection and Diagnosis (MDD) research [1, 2, 3, 4]. Phonemes capture pronunciation details that letters cannot, removing ambiguity and making text computationally interpretable.

The Qur'an is recited predominantly in the Hafs 'an Asim style (*riwaya*) [5, 6], and introduces unique linguistic complexity through Tajweed rules. These rules govern recitation, yet many are not explicitly marked in the Qur'anic script and can only be inferred by knowledgeable reciters.

Applications of a Qur'anic phonemizer include:

- **ASR / MDD:** Using phonemes enables models to detect specific mispronunciations and provide targeted feedback—a standard practice in fine-tuning pretrained transformer models.
- **Text-to-Speech (TTS):** While synthetic Qur'anic recitations are unnecessary due to abundant professional recordings, controlled synthesis can deliberately alter Tajweed or diacritics to generate mispronounced data, supporting augmentation for ASR and MDD training [7].
- **Linguistic analysis:** Rhyme structures, phoneme frequencies, and Tajweed rule distributions, offering insights into the miraculous stylistic and melodic properties of the Qur'an.

Given the lack of such a tool, we address these needs by presenting an open-source Qur'anic phonemizer tailored to Hafs. This tool opens new directions for speech research, pedagogy, and Muslim-focused ML, bringing Muslims closer to the spirituality of reciting the Qur'an.

Phonemizers have been developed for many languages, including vowelised Modern Standard Arabic (MSA) by Halabi and Wald [8]. However, these do not capture the intricacies of Tajweed, nor do they handle unique Qur'anic symbols absent in MSA text.

Many studies have analysed Qur'anic phonetics linguistically [9, 10, 11, 12, 13, 14, 15]. Some have proposed computational rule-based phoneme transcription and text preprocessing frameworks

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: MusIML.

[16, 17, 18, 19, 20], however none comprehensively modelled all Tajweed rules and script-specific symbols nor provided a unified open-source grapheme-to-phoneme tool.

Datasets for Qur'anic recitation are scarce, especially with phoneme labels. Very recently, the Iqra'Eval shared task [7] augmented CommonVoice [21] Arabic data with Qur'anic recitations read without Tajweed, labelling phonemes with the MSA phonemizer of Halabi, and also released the first benchmark phoneme-labelled test set, QuranMB.v1, with mispronunciation samples.

Our phonemizer complements such efforts by providing an automatic method to generate phoneme labels. Given correctly recited audio and their verse references, it can label at scale, enabling large Qur'anic datasets such as EveryAyah [22] to be phonemized for many reciters.

## 2 Methodology

### 2.1 High-level approach

The phonemizer uses a modular object-oriented design, with context-aware classes for letters and words-essential since Tajweed rules depend on context. Specific letter classes override base behaviour to define rule-specific phonemization.

We use the Qur'anic Uthmani script from QUL [23]: a JSON file of words with metadata, enabling the phonemizer to accept flexible references and outputting International Phonetic Alphabet (IPA) [24] phonemes with additional symbols for Tajweed, yielding 71 phonemes: 28 consonants, 24 geminated consonants (*shaddah*), 8 vowels, and 11 Tajweed phonemes. A full mapping of graphemes to phonemes is provided in Appendix A.1.

### 2.2 Letters

Most letters map directly to their base phoneme. Special handling is required for context-sensitive letters such as *meem*, *noon*, *lam*, *raa* and vowels. We also resolve specific ambiguities: 1) *taa marbuta* is converted to *taa* in continuation, or *haa* when stopping; 2) *yaa/waw* are long vowels if undiacritised and preceded by kasra/damma respectively, otherwise consonantal; 3) *alef maksura* is treated as consonant *yaa*, vowel *yaa*, or vowel *alef* based on context; 4) all *hamza* variants map to the same phoneme; 5) *hamzat-al wasl* is either silent or pronounced with a diacritic determined from context.

### 2.3 Shaddah

Shaddah is represented as doubling of the base phoneme. For example, /b/ with shaddah becomes /bb/. Notably, *ghayn* and *hamza* never appear geminated in the Qur'an.

### 2.4 Vowels

Short vowels (fatha, damma, kasra) are modelled as /a/, /u/, /i/, with long versions /a:/, /u:/, /i:/. Fatha and alef are conditionally emphatic depending on the preceding consonant.[1]

### 2.5 Special words

Certain words in the Qur'an are unique and cannot be derived purely from orthography (Appendix A.2). For example, *alif lam meem* is //aliffla:mi:m// instead of //alm//.

### 2.6 Pausing

Pausing (*waqf*) alters realisation of phonemes. The phonemizer automatically applies start- and stop-word rules, with the ability to specify additional optional stop signs.

**Starting words** 1) shaddah is removed (no word may begin with shaddah); 2) cross-word Tajweed rules (Idgham, Iqlab, Ikhfaa) are not applied.

[1] *kha*, *sad*, *dad*, *ghayn*, *tta*, *qaf*, *ttha*, emphatic *lam* and *raa*.

**Stopping words** 1) hamza + fathatan is converted to hamza + alef; 2) taa marbuta is converted to haa; 3) tanween is removed for alef/alef maksura; 4) last letter diacritics are dropped; 5) cross-word Tajweed rules (Idgham, Iqlab, Ikhfaa) are not applied; 6) Qalqala is applied to compatible letters. [2]

### 2.7 Tajweed rules

The phonemizer encodes all acoustic Tajweed rules (see Table 1) except for elongation (*Madd*). This is because Madd only affects the duration of a vowel but phonetically represents the same sound. Elongated vowels are mapped to their corresponding long-vowel symbol, regardless of Madd type. For detailed explanations and conditions of the Tajweed rules, see Appendix A.3.

Table 1: Tajweed rules examples. Phoneme(s) affected by the Tajweed rule are bolded for clarity.

| Rule | Text Example | Phonemes |
|---|---|---|
| Ghunnah | ثُمَّ إِنَّكُم | // θu**m̃**a ʔi**ñ**akum // |
| Idgham (Ghunnah) | خَلَقَكُم مِّن نَّفْسٍ وَاحِدَة | // xaˤlaqaˤku**m̃** i **ñ**afsi **w̃**a:ħidah // |
| Idgham (no Ghunnah) | كُلِّ شَيْءٍ رِّزْقًا مِّن لَّدُنَّا | // kulli ʃajʔ**i rr**izqa m̃**i ll**aduña: // |
| Idgham (other) | إِن كِدتَّ لَتُرْدِين | // ʔiŋ ki**tt**a laturˤdi:n // |
| Ikhfaa | سَبحًا طَوِيلا | // sabQħa**ŋ**ˤ tˤaˤwi:la: // |
| Ikhfaa Shafawi | أَمَدَّكُم بِأَنْعَام | // ʔamaddaku**ŋ** biʔanˤa:m // |
| Iqlab | لَا يَنبَغِي لِأَحَدٍ مِّن بَعْدِي | // la: ja**m̃**baɣi: liʔaħadi m̃i**m̃** baˤdi: // |
| Qalqala | لَمْ يَلِدْ وَلَمْ يُولَدْ | // lam jalid**Q** walam ju:lad**QQ** // |
| Hamzat al-Wasl | آلْجَوَارِ آلْكُنَّس | // **ʔa**lʒawa:r**i l**kuñas // |
| Iltiqaa as-Sakinayn (vowels) | آهْدِنَا آلصِّرَاطَ آلْمُسْتَقِيم | // ʔihdin**a sˤsˤ**irˤaˤ:tˤ**aˤ l**mustaqi:m // |
| Iltiqaa as-Sakinayn (tanween) | وَنَادَى نُوحٌ آبْنَه | // wana:da: nu:ħun**i** bQnah // |
| Tafkheem/Tarqeeq Lam | وَتَوَكَّلْ عَلَى اللَّه | // watawakkal ʕala **lˤlˤaˤ:**h // |
| Tafkheem/Tarqeeq Raa | فَقَدَرَ عَلَيْهِ رِزْقَه | // faqaˤda**rˤaˤ** ʕalajhi **r**izqaˤh // |

We note that Ikhfaa and Qalqala phonemes occur in two variants (emphatic or normal, sughra or kubra respectively). While the phonemizer outputs all variants by default, depending on the application, it may be useful to merge them as one base phoneme, as the phonetic/acoustic differences are minor.

## 3 Results

Phonemizer outputs were verified through systematic manual inspection by researchers familiar with Qur'anic recitation and phonetic rules. Multiple chapters spanning all Tajweed rules and special orthographic cases were examined. Detected inconsistencies were iteratively corrected until stable outputs were achieved across all evaluated chapters.

To complement this internal verification, we further validated the phonemizer externally in our other work [25] by training a transformer model using its generated phoneme labels and evaluating them on benchmark datasets. The resulting model achieved 3.9% and 2.6% phoneme error rate (PER) on the Iqra'Eval Qur'an and professional reciters' development sets, respectively, and 84.9% accuracy on the QuranMB test set, demonstrating that the phonemizer's outputs align with expert-labelled benchmarks and can generalise effectively in downstream speech tasks.

[2] *qaf*, *tta*, *baa*, *jeem*, *dal*.

That said, as there is currently no established automatic evaluation protocol for Qur'anic phonemization, we invite the community to further verify outputs and report any discrepancies to strengthen and refine the tool.

Figure 1 shows the phoneme distribution of the entire Qur'an. [3] Vowels dominate the distribution, with /a/ (fatha) contributing just over 17.5%. Consonants like /l/ and /n/ are very common, and the least frequent phonemes are mostly shaddah phonemes given their scarcity in Qur'anic words. Additional emphatic Lam and categorical distributions can be found in Appendix A.4.

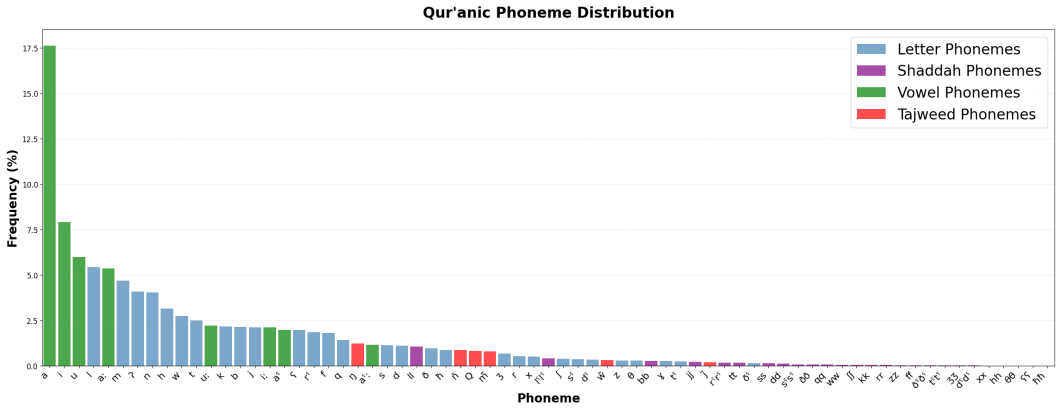


Figure 1: Phoneme distribution across the Qur'an.

## 4 Conclusion and limitations

In this work we presented an open-source Qur'anic phonemizer tailored to the Hafs riwaya, modelling the full set of Tajweed rules and Qur'anic symbols. Its modular design, context-aware architecture, and comprehensive phoneme inventory enable accurate text-to-phoneme conversion and open new directions for ASR, MDD, linguistic analysis, and educational applications.

Currently, the phonemizer only supports the Hafs riwaya. Other recitation traditions use distinct Tajweed rules, orthographic conventions, and symbol sets. Although the modular architecture allows straightforward extension in principle, such adaptation requires access to a digital script and specialised linguistic expertise. We view this as a key avenue for future work and invite collaboration to expand the phonemizer to additional riwayat.

Another limitation concerns the handling of edge cases where words admit multiple valid recitations—such as variations involving *seen*/*sad*, or *fatha*/*damma*. While these exceptions are few, the phonemizer currently produces only one canonical output based on the default orthography in the Uthmani script. Incorporating alternative realisations could enhance its value for constructing pronunciation dictionaries that represent all valid phoneme variants.

Finally, while the system is intended for research and pedagogical use, responsible application is essential given it concerns the sacred text of the Qur'an. Synthetic recitation generation or modification should always respect religious boundaries and avoid misuse such as imitating or replacing authentic recitation. To safeguard ethical use, we encourage open licensing, clear usage guidelines, and collaboration with scholars to ensure spiritual integrity in downstream applications.

Overall, this work establishes a strong foundation for future development, and we hope it catalyses research at the intersection of Qur'anic recitation, speech technology, education, and ethical use.

[3] Phonemizing the entire Qur'an takes <10 seconds on a normal CPU. To use the phonemizer efficiently, it is recommended to phonemize a range e.g., "1-114" and split by "verse_sep" rather than individual verses in a loop.

## References


[1] Wai-Kim Leung, Xunying Liu, and Helen Meng. CNN-RNN-CTC-based end-to-end mispronunciation detection and diagnosis. In *Proceedings of the 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8132–8136, Brighton, United Kingdom, May 2019. doi: 10.1109/icassp.2019.8682654. URL https://doi.org/10.1109/icassp.2019.8682654.

[2] Yiqing Feng, Guanyu Fu, Qingcai Chen, and Kai Chen. SED-MDD: Towards sentence dependent end-to-end mispronunciation detection and diagnosis. In *Proceedings of the 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3492–3496, Barcelona, Spain, May 2020. doi: 10.1109/icassp40776.2020.9052975. URL https://doi.org/10.1109/icassp40776.2020.9052975.

[3] Bi-Cheng Yan, Meng-Che Wu, Hsiao-Tsung Hung, and Berlin Chen. An end-to-end mispronunciation detection system for L2 English speech leveraging novel anti-phone modeling, 2020. URL https://arxiv.org/abs/2005.11950.

[4] S. M. Witt and S. J. Young. Phone-level pronunciation scoring and assessment for interactive language learning. *Speech Communication*, 30(2–3):95–108, 2000. doi: 10.1016/S0167-6393(99)00044-8. URL https://doi.org/10.1016/S0167-6393(99)00044-8.

[5] Adrian Alan Brockett. *Studies in Two Transmissions of the Qur'an*. PhD thesis, University of St Andrews, St Andrews, UK, 1985. URL https://hdl.handle.net/10023/2770.

[6] Philipp Bruckmayr. Challenging the cairo edition: The king fahd Qur'an complex, its medina Qur'an and its translations. *MIDÉO (Mélanges de l'Institut dominicain d'études orientales)*, 39:211–257, 2024. URL https://journals.openedition.org/mideo/9400.

[7] Yassine El Kheir, Omnia Ibrahim, Amit Meghanani, Nada Almarwani, Hawau Olamide Toyin, Sadeen Alharbi, Modar Alfadly, Lamya Alkanhal, Ibrahim Selim, Shehab Elbatal, Salima Mdhaffar, Thomas Hain, Yasser Hifny, Mostafa Shahin, and Ahmed Ali. Towards a unified benchmark for arabic pronunciation assessment: Qur'anic recitation as case study, 2025. URL https://arxiv.org/abs/2506.07722.

[8] Nawar Halabi and Mike Wald. Phonetic inventory for an Arabic speech corpus. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Helene Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 734–738, Portorož, Slovenia, May 2016. European Language Resources Association. URL https://aclanthology.org/L16-1116/.

[9] Fatima Abdullah Almousa and Faisal M. Al-Mohanna. The Qur'anic conditionally pharyngealized sounds: An optimality theory perspective. *Arab World English Journal for Translation and Literary Studies*, 5 (3):125–150, August 2021. doi: 10.24093/awejtls/vol5no3.10. URL https://doi.org/10.24093/awejtls/vol5no3.10.

[10] S. S. S. Alsurf. *The Phonetics of the Qur'anic Pharyngealised Sounds: Acoustic and Articulatory Studies*. Phd thesis, Macquarie University, Sydney, Australia, 2013. Doctoral dissertation.

[11] Marijn van Putten. Inferring the phonetics of Qur'anic Arabic from the Qur'anic consonantal text. *International Journal of Arabic Linguistics*, 5(1):1–19, 2019.

[12] S. A. Al-Hashmi. The phonology of nasal *n* in the language of the holy Qur'an. Master's thesis, University of Victoria, 2004.

[13] Abbas Jawdat Rahim and Maryam Falih Ahmad. The phenomenon of qalqala in Qur'an recitation. *International Journal of Business and Social Science*, 10(7):145–157, July 2019. doi: 10.30845/ijbss.v10n7p16. URL https://doi.org/10.30845/ijbss.v10n7p16.

[14] Tareq Altalmas, Salmiah Ahmad, Wahju Sediono, and Surul Shahbudin Hassan. Qur'anic letter pronunciation analysis based on spectrogram technique: Case study on qalqalah letters. In *Proceedings of the 11th International Conference on Artificial Intelligence Applications and Innovations (AIAI 2015): Special Tracks and Workshops*, volume 1539 of *CEUR Workshop Proceedings*, pages 14–22, 2015.

[15] Husain Nasir, Achmad Abubakar, Muhammad Galib, Amrah Kasim, and Muhsin Mahfudz. Phonology and vowel sounds in the Qur'an: Perspectives on classical and modern phonetic rules. *ELOQUENCE: Journal of Foreign Language*, 2(2):106–123, August 2023. doi: 10.58194/eloquence.v2i2.655. URL https://doi.org/10.58194/eloquence.v2i2.655.

[16] Nadjla Bettayeb and Mhania Guerti. Study to build a holy Qur'an text-to-speech system. *International Journal on Islamic Applications in Computer Science and Technology*, 7:1–10, December 2019.

[17] Clare Brierley, Majdi Sawalha, Barry Heselwood, and Eric Atwell. A verified Arabic-IPA mapping for Arabic transcription technology, informed by Qur'anic recitation, traditional Arabic linguistics, and modern phonetics. *Journal of Semitic Studies*, 61:157–186, 2016. doi: 10.1093/jss/fgv035. URL https://doi.org/10.1093/jss/fgv035.

[18] Sameh Bellegdi and Husni Al-Muhtaseb. Automatic rule-based phonetic transcription and syllabification for Qur'anic text. *International Journal on Islamic Applications in Computer Science and Technology*, 3: 17–28, December 2015.

[19] Majdi Sawalha, Clare Brierley, Eric Atwell, and James Dickins. Text analytics and transcription technology for Qur'anic Arabic. *International Journal on Islamic Applications in Computer Science and Technology*, 5, June 2017.

[20] Eric Atwell, Majdi Sawalha, and Clare Brierley. Automatically generated, phonemic Arabic-IPA pronunciation tiers for the boundary annotated Qur'an dataset for machine learning (version 2.0). Dataset release, January 2014. URL https://doi.org/10.13140/2.1.2887.2640.

[21] Rosana Ardila, Megan Branson, Kelly Davis, Michael Henretty, Michael Kohler, Josh Meyer, Reuben Morais, Lindsay Saunders, Francis M. Tyers, and Gregor Weber. Common voice: A massively-multilingual speech corpus, 2019. URL https://arxiv.org/abs/1912.06670.

[22] Anonymous. Everyayah dataset. https://everyayah.com/, 2010. Online.

[23] Tarteel AI. Qur'an script QPC hafs – word by word. https://qul.tarteel.ai/resources/quran-script/312. Online.

[24] International Phonetic Association. *Handbook of the International Phonetic Association: A Guide to the Use of the International Phonetic Alphabet*. Cambridge University Press, Cambridge, 1999. doi: 10.1017/9780511807954. URL https://doi.org/10.1017/9780511807954.

[25] Ahmed Ibrahim. Hafs2Vec: A system for the IqraEval Arabic and qur'anic phoneme-level pronunciation assessment. In Kareem Darwish, Ahmed Ali, Ibrahim Abu Farha, Samia Touileb, Imed Zitouni, Ahmed Abdelali, Sharefah Al-Ghamdi, Sakhar Alkhereyf, Wajdi Zaghouani, Salam Khalifa, Badr AlKhamissi, Rawan Almatham, Injy Hamed, Zaid Alyafeai, Areeb Alowisheq, Go Inoue, Khalil Mrini, and Waad Alshammari, editors, *Proceedings of The Third Arabic Natural Language Processing Conference: Shared Tasks*, pages 453–456, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-356-2. URL https://aclanthology.org/2025.arabicnlp-sharedtasks.62/.

# A Technical Appendices and Supplementary Material

## A.1 Phoneme inventory

Table 2: Consonant letters and phonemes (gemination shown as doubled).

| Letter | Phoneme | Letter | Phoneme |
|---|---|---|---|
| أ ، إ ، ء ، ؤ ، ئ | ʔ | ب | b / bb |
| ت | t / tt | ث | θ / θθ |
| ج | ʒ / ʒʒ | ح | ħ / ħħ |
| خ | x / xx | د | d / dd |
| ذ | ð / ðð | ر | r / rˤ / rr / rˤrˤ |
| ز | z / zz | س | s / ss |
| ش | ʃ / ʃʃ | ص | sˤ / sˤsˤ |
| ض | dˤ / dˤdˤ | ط | tˤ / tˤtˤ |
| ظ | ðˤ / ðˤðˤ | ع | ʕ / ʕʕ |
| غ | ɣ | ف | f / ff |
| ق | q / qq | ك | k / kk |
| ل | l / ll / lˤlˤ | م | m |
| ن | n | هـ | h / hh |
| و | w / ww | ى ، ي | j / jj |

Table 3: Vowels phonemes.

| Symbol | Phoneme |
|---|---|
| ◌َ | a / aˤ |
| ◌ُ | u |
| ◌ِ | i |
| ا ، ى | a: / aˤ: |
| و | u: |
| ى ، ي | i: |

Table 4: Tajweed rule phonemes.

| Rule | Phoneme |
|---|---|
| Iqlab | m̃ |
| Idgham | ñ / m̃ / j̃ / w̃ |
| Ikhfaa | ŋ (Light/Shafawi) ŋˤ (Heavy) |
| Qalqala | Q (Sughra) QQ (Kubra) |
| Tafkheem | lˤlˤ (Lam in “Allah”) rˤ / rˤrˤ (Raa) |

### A.2 Special words

Table 5: Special Qur'anic words with phoneme sequences.

| Word | Phonemes | Word | Phonemes |
|---|---|---|---|
| حم | ħa:mi:m | الم | aliffla:mi:m |
| الر | aliffla:mrˤaˤ: | المص | aliffla:mi:msˤaˤ:dQ |
| المر | aliffla:mi:mrˤaˤ: | كهيعص | ka:ffha:ja:ʕajɲsˤaˤ:dQ |
| طسم | tˤaˤ:si:mi:m | طس | tˤaˤ:si:n |
| طه | tˤaˤ:ha: | يس | ja:si:n |
| ص | sˤaˤ:dQ | عسق | ʕajɲsi:ŋqaˤ:ff |
| ق | qaˤ:ff | ن | nu:n |
| مجراها | maʒQri:ha: | نُـنجِي | nuŋʒi |

### A.3 Tajweed rule details

**Ikhfaa**

- Applies within or across words.
- Trigger: *noon sakinah* or *tanween* followed by one of the 15 letters: ت، ث، ج، د، ذ، ز، س، ش، ص، ض، ط، ظ، ف، ق، ك.
- Heavy Ikhfaa: if the following letter is one of the emphatic set خ، ص، ض، غ، ط، ق، ظ.

**Ikhfaa Shafawi**

- Cross-word only.
- Trigger: *meem sakinah* followed by ب.
- Realised with the same nasalised Ikhfaa phoneme.

**Iqlab**

- Applies within or across words.
- Trigger: *noon sakinah* or *tanween* followed by ب.
- Transformation: noon is converted to *meem* and nasalised.

**Idgham**

**With Ghunnah**

- Cross-word only.
- Trigger: *noon sakinah* or *tanween* followed by one of ي، ن، م، و.
- Effect: sounds merge and become nasalised.

**Without Ghunnah**

- Cross-word only.
- Trigger: *noon sakinah* or *tanween* followed by ل or ر.
- Effect: noon becomes silent.

**Lam Shamsiyah**

- In ال, if ل is followed by a sun-letter, the lam is silent.

**Other Forms**

- *Mutamathilayn*, *Mutaqaribayn*, *Mutajanisayn*.

- General pattern: a letter without sukoon followed by the same or a similar-sounding letter → the first is silent.

**Qalqala**

- Trigger: one of the qalqala letters with sukoon ق، ط، ب، ج، د.
- Realization: pronounced with an echo /Q/ phoneme.
- Types:
  - *Sughra*: in the middle of a word or final letter of a continuing word.
  - *Kubra*: final letter of a pausing word.

**Hamzat-al Wasl**

- Silent if in the middle of a word, or at the start during continuous recitation.
- Pronounced if starting on that word:
  - If a noun starting with ال, hamza takes fatha.
  - If a verb and the third letter has damma, hamza takes damma; otherwise kasra.
- *Iltiqa al-sakinayn* (meeting of two sakin letters):
  - If the previous letter is a long vowel, it is shortened (e.g. /a:/→a).
  - If the previous letter has tanween, insert kasra /i/ after the tanween.

**Lam Tafkheem/Tarqeeq**

- Occurs only in Allah word patterns.
- If preceded by fatha or damma → emphatic (heavy) lam.
- Otherwise → light lam.

### A.4 Additional results and distributions

Table 6: Allah word patterns: totals and heavy/light lam counts.

| Word | Total | Heavy | Light |
|---|---|---|---|
| اللَّهُم | 5 | 3 | 2 |
| اللَّه | 2153 | 1731 | 422 |
| ءَاللَّه | 2 | 2 | |
| وَاللَّه | 240 | 240 | |
| فَاللَّه | 6 | 6 | |
| تَاللَّه | 8 | 8 | |
| وَتَاللَّه | 1 | 1 | |
| لِلَّه | 116 | | 116 |
| وَلِلَّه | 27 | | 27 |
| فَلِلَّه | 6 | | 6 |
| باللَّه | 139 | | 139 |
| أَبِاللَّه | 1 | | 1 |
| **Total** | **2704** | **1991** | **713** |

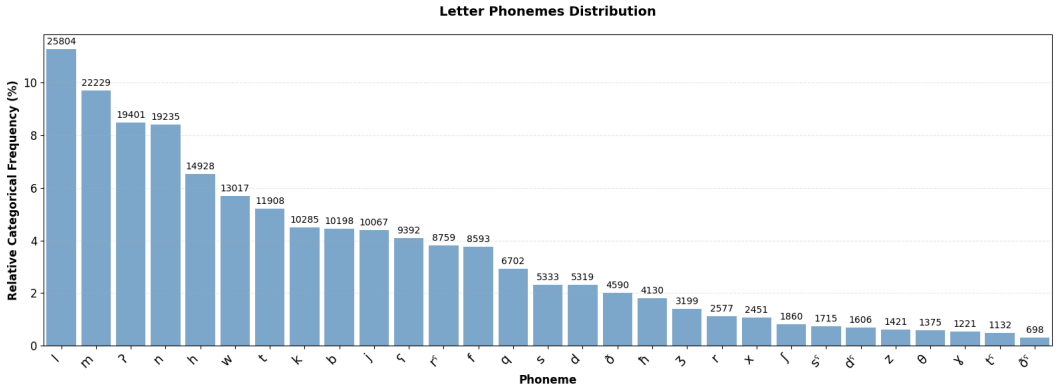


Figure 2: Phoneme distribution of letter phonemes.

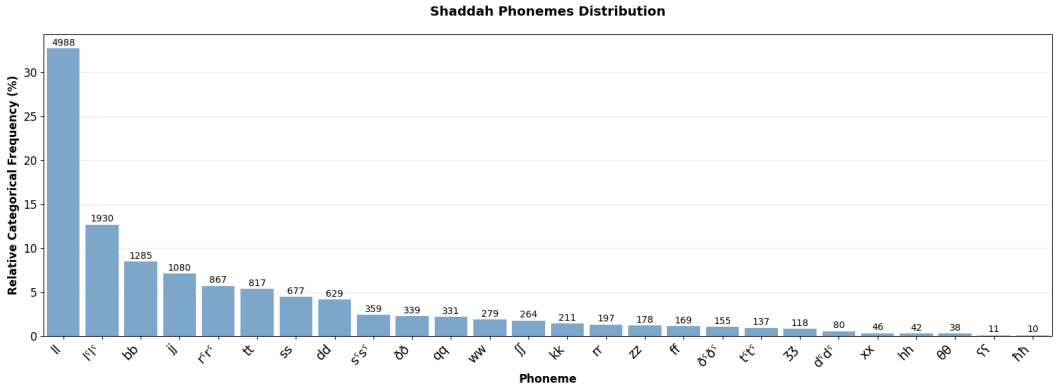


Figure 3: Phoneme distribution of shaddah phonemes.

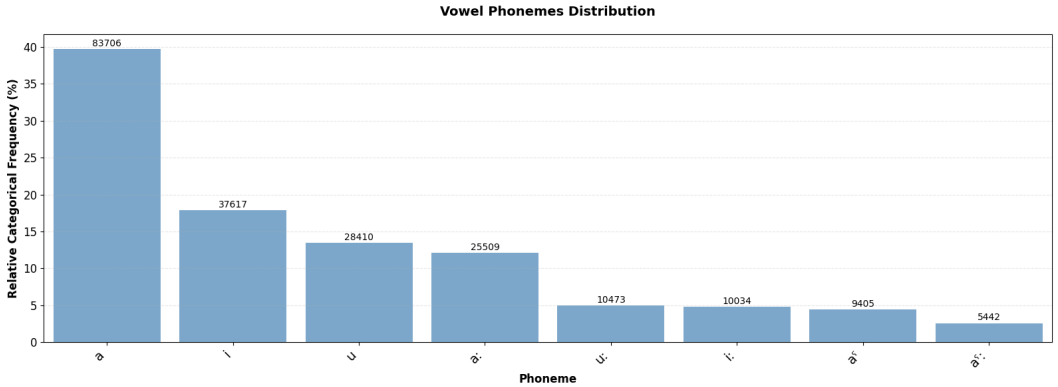


Figure 4: Phoneme distribution of vowel phonemes.

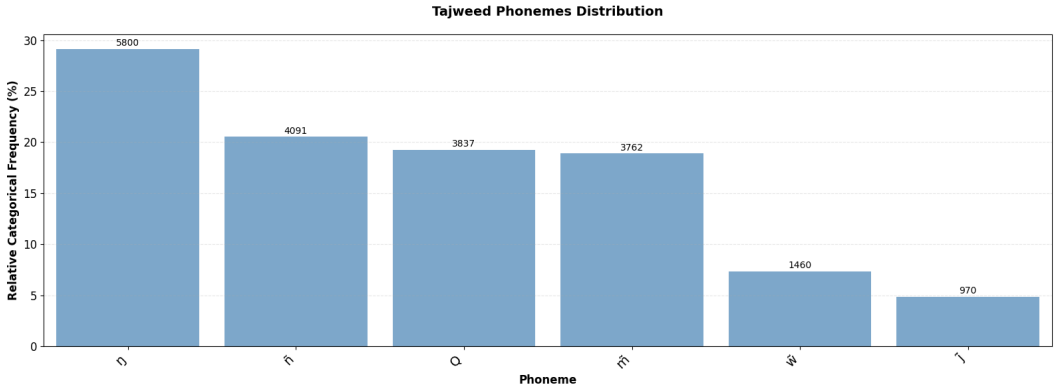


Figure 5: Phoneme distribution of Tajweed phonemes.

## NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction clearly state the contributions (a Qur'anic phonemizer tailored to Hafs with full Tajweed support, modular design, phoneme inventory) and scope (limited to Hafs). These match the results and limitations discussed later.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: A dedicated Limitations section discusses constraints: Hafs-only support, handling of edge cases, and canonical-only outputs. Future directions are also outlined.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.

- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

   Justification: The paper does not present formal theoretical results, theorems, or proofs. It is a system paper focused on design and empirical validation.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The methodology section describes architecture, phoneme inventory, rule handling, and evaluation procedure. Outputs are verified systematically. Together with open-source code, this allows reproduction. Detailed rules and instructions are provided in the supplementary material.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case

of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.

- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The phonemizer Github code and user interface are explicitly released as open source, with links (redacted in anonymised version). This provides instructions for usage and reproduction.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: While no ML training is run, the paper specifies details of implementation: modular design, phoneme inventory, rules encoded, handling of letters, vowels, pausing, and Tajweed rules.

   Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [NA]

   Justification: No statistical experiments are reported. Verification is manual and qualitative, not quantitative.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
   - If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: No heavy ML experiments are conducted. The phonemizer is computationally light (entire Qur'an <10s on CPU).

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The work aligns with NeurIPS Code of Ethics: Qur'anic script and rules are open-access, no privacy or fairness risks are implicated.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Positive impacts are outlined (ASR/MDD, pedagogy, linguistic analysis, speech research). Negative misuse (e.g., potential misuse for automatic recitation replacement or misrepresentation of Tajweed) is not discussed in depth, but limitations restrict such risks.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: The released code poses low misuse risk compared to large pretrained models. It does not generate content, only rule-based phoneme sequences.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: The Qur'anic script source QUL is credited, and datasets like EveryAyah and Iqra'Eval are cited. Their licenses and usage terms are respected.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: The phonemizer is a new open-source asset. Documentation is provided in the paper and detailed in the GitHub repository.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: No crowdsourcing or human-subject studies are conducted.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: Not applicable, as there are no human-subject experiments.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification: No LLMs are used in the core methodology.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.