TajweedAI: A Hybrid ASR-Classifier for Real-Time Qalqalah Detection in Quranic Recitation

Nabhan Mazid
Rutgers University
New Brunswick, NJ
nm1088@scarletmail.rutgers.edu

Muaz Ahmad
Rutgers University
New Brunswick, NJ
ma1962@scarletmail.rutgers.edu

Abstract

Proper recitation of the Holy Quran is governed by a complex set of phonetic rules known as Tajweed, where minor pronunciation errors can significantly alter meaning. While modern Artificial Intelligence (AI) tools excel at transcription, they largely lack the capability to provide corrective feedback on pronunciation quality. This paper introduces TajweedAI, a novel system designed to bridge this gap by offering real-time, fine-grained phonetic analysis for Quranic learners. We present a hybrid architecture that combines a state-of-the-art Automatic Speech Recognition (ASR) model for temporal alignment with a dedicated binary classifier for phonetic rule verification. As a case study, we focus on the acoustically complex Tajweed rule of *Qalqalah*—the characteristic "echoing" of specific plosive consonants. This paper details an iterative experimental methodology, beginning with a baseline model achieving 58.33% accuracy and culminating in a highly specialized classifier trained via hard negative mining. This final model achieved 100% accuracy on its specialized internal validation set for the challenging case of the word *al-Falaq*. However, a limited external evaluation indicated challenges in generalization, yielding 57.14% accuracy. This work validates a scalable framework for automated Tajweed correction, presenting a significant step for Computer-Assisted Pronunciation Training (CAPT) in Quranic studies.

1 Introduction

The recitation of the Holy Quran is a foundational practice in Islam, governed by *Tajweed*, a comprehensive set of rules for correct pronunciation. Adherence to these rules is essential for preserving the divine text's intended meaning, as even subtle phonetic errors can alter a word's definition (3; 4). Traditionally, mastering Tajweed requires years of one-on-one instruction, but access to expert teachers is a significant challenge for millions, creating a need for scalable learning tools (2).

While AI-powered applications have revolutionized Quranic engagement with features like real-time transcription and memorization assistance, their core competency lies in recognizing *what* is said, not *how* it is said. ASR systems are designed to be robust to the very phonetic and prosodic variations that Tajweed regulates, making them inherently unsuitable for providing the corrective feedback necessary for pronunciation training (5). This distinction between transcription and pronunciation analysis represents the central gap our work aims to address.

This paper introduces TajweedAI, a novel Computer-Assisted Pronunciation Training (CAPT) system for Quranic Arabic. We focus on a single, acoustically complex rule, *Qalqalah*, as a robust test case. Our principal contributions are: 1) A hybrid ASR-classifier architecture that leverages a pre-trained ASR model for temporal localization followed by a dedicated, lightweight acoustic classifier for phonetic evaluation. 2) The successful development of a high-fidelity *Qalqalah* classifier that achieves

100% test accuracy on a challenging, specific word through a targeted training strategy. 3) An iterative experimental framework, including the novel application of hard negative mining, which serves as a blueprint for developing classifiers for other complex phonetic rules.

2 System Architecture and Design

The TajweedAI system is an end-to-end pipeline designed to progress from data collection to real-time user feedback. Its architecture consists of a data preprocessing stage and a real-time analysis loop.

2.1 Data Acquisition and Preprocessing

A high-quality dataset is the foundation of the system. Audio data was sourced from Tarteel's Quranic Universal Library (QUL), which provides high-quality, verse-segmented recordings from numerous world-renowned reciters. To train our models, precise word-level timestamps were required. These were generated using a forced alignment tool ('ctc-forced-aligner') that leverages a state-of-the-art Arabic ASR model ('nvidia/stt-ar-fastconformer.hybrid-large.pcd-v1.0') to align the audio with a ground-truth transcript (6). This approach provided exceptionally accurate word segments, bypassing the need to train a custom aligner.

To generate a pristine labeled dataset for initial experiments, we developed a custom GUI tool, the "Qalqalah Annotator" (Figure 1). This tool allowed an expert user to visually inspect a waveform, listen to the audio segment, and precisely mark the start and end times of the *Qalqalah* "bounce" event, creating a small but high-confidence dataset of positive examples.

2.2 Real-Time Feedback Loop

The user-facing application operates through a multi-step, real-time feedback loop:

1. **Recording Transcription:** The user records their recitation of a specific verse (*ayah*). This audio is passed to the NVIDIA ASR model to get a transcript and identify candidate words where a *Qalqalah* rule applies.
2. **Segmentation:** Once a candidate word is identified, the forced aligner determines its precise start and end timestamps within the user's audio.
3. **Extraction Classification:** The audio segment for the target word is isolated. This slice is transformed into a set of acoustic features (e.g., MFCCs, spectral centroid, zero-crossing rate) and passed to our trained binary classifier.
4. **Feedback:** The classifier returns a binary output ("Qalqalah" or "No Qalqalah"), which is rendered in the UI to provide immediate, specific feedback on the user's pronunciation. The UI is shown in the Appendix (Figure 2).

3 Experimental Evaluation

The development of an effective *Qalqalah* detector was an iterative process over three distinct experiments. Audio features were extracted for all experiments, including Mel-Frequency Cepstral Coefficients (MFCCs), which are highly effective at capturing the phonetic qualities of speech (7).

Experiment 1: Baseline Classification. The initial experiment used a small, meticulously curated dataset of 56 samples (28 positive, 28 negative). An SVM classifier trained on this data achieved a test accuracy of only **58.33%**, marginally better than random chance.

Experiment 2: Scaling with a Mass-Generated Dataset. A larger, more "noisy" dataset of 301 samples (70 positive, 231 negative) was automatically generated. A Random Forest model achieved a higher test accuracy of **83.6%**. However, its recall for the "Qalqalah" class was only 0.500, indicating bias.

Experiment 3: High-Precision Detection via Hard Negative Mining. The insights from the first two experiments led to a highly targeted third experiment using hard negative mining, focusing exclusively on the word *al-falaq*. This strategy forces the model to learn the single acoustic feature separating the classes. Both SVM and Random Forest classifiers trained on this specialized dataset

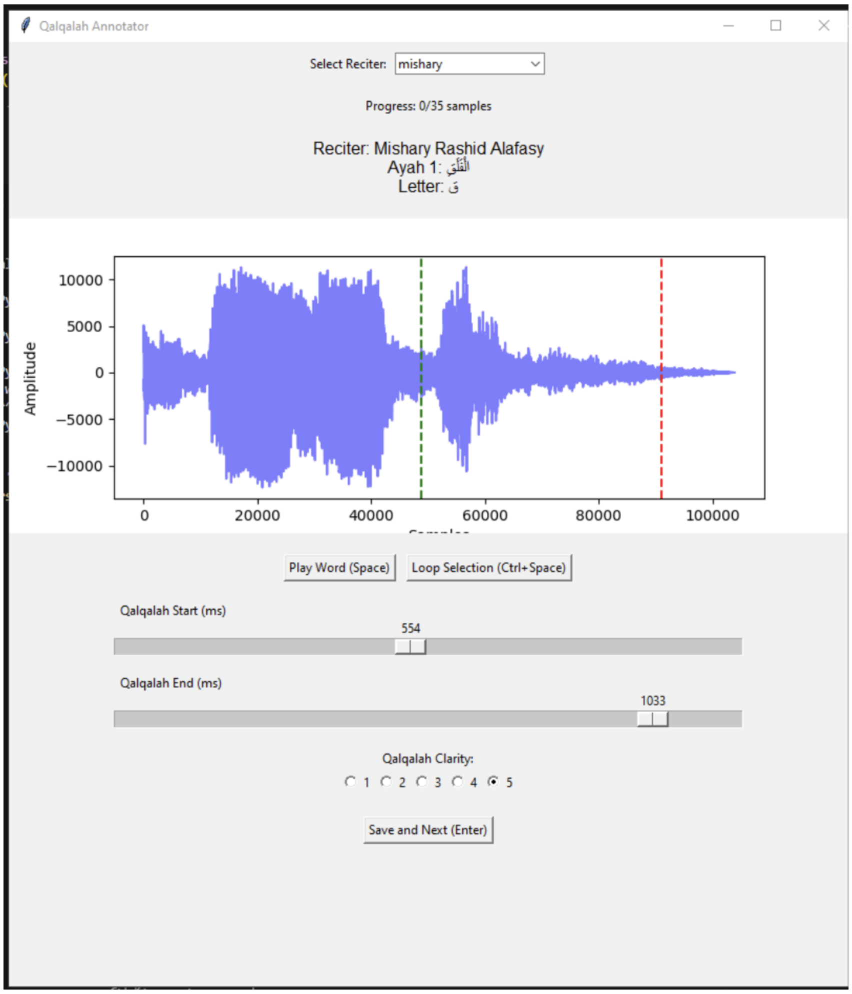

Figure 1: The Qalqalah Annotator GUI, used for creating the ground-truth dataset by allowing an expert to mark the precise start and end times of the phonetic event.

achieved a perfect test accuracy of **100.0%** on the internal validation set. This result provides compelling evidence that for fine-grained phonetic error detection, the acoustic specificity of training data is far more critical than raw quantity. A summary is in Table 1.

Table 1: Comparative Performance Across Three Experimental Stages

Experiment	Methodology	Test Accuracy	Recall (Qalqalah)
1: Baseline	Small, Curated (n=56)	58.33%	0.580
2: Scaled	Large, Noisy (n=301)	83.6%	0.500
3: Targeted	Hard Negative Mining (n=29)	**100.0%**	**1.000**

4 Related Work

Our work is situated at the intersection of automated Tajweed rule detection, Arabic CAPT, and deep learning for Quranic speech. Early research employed traditional models like SVMs and HMMs (1; 8). More recent work has shifted towards deep learning to classify rules such as *Madd* and *Ghunnah* (2; 9). Broader Arabic CAPT has focused on detecting errors by non-native learners using Goodness of Pronunciation (GOP) scores (10). Our work contributes by proposing a hybrid architecture that decouples word localization from fine-grained phonetic verification for the specific rule of *Qalqalah*.

5 Conclusion

This paper presented TajweedAI, a system designed to provide real-time, corrective phonetic feedback for Quranic recitation learners. We designed and evaluated a hybrid architecture using a state-of-the-art ASR model and a specialized acoustic classifier. We demonstrated that a data strategy of hard negative mining is exceptionally effective for developing a high-precision phonetic error detector, achieving 100% accuracy on its internal validation set.

However, initial external testing yielded 57.14% accuracy, underscoring the challenge of generalization from limited data. Despite this limitation, primarily attributed to the constraints of a project timeline, this work validates a highly promising and scalable methodology for moving beyond simple transcription towards meaningful pronunciation analysis, laying a crucial foundation for future advancements in Quranic CAPT.

6 Future Work

The successful development of this proof-of-concept opens several exciting avenues for future work:

- **Expansion to All Qalqalah Letters:** The immediate next step is to apply the successful hard negative mining methodology to the remaining four *Qalqalah* letters to build a comprehensive detector for the rule.
- **Addressing Other Tajweed Rules:** The hybrid ASR-classifier framework can be extended to other acoustically-defined rules, such as *Madd* (correct vowel prolongation) and *Ghunnah* (nasalization), which are common challenges for learners.
- **Improving Feedback Granularity:** Moving beyond binary ”correct/incorrect” feedback is crucial for a real learning tool. Future versions could classify the type of error (e.g., ”Qalqalah too weak,” ”Qalqalah over-exaggerated into a full vowel”).
- **Open-Sourcing Contributions:** As envisioned in the project’s initial planning, we aim to open-source the Qalqalah Annotator tool and the curated datasets as a valuable contribution to the research community, fostering further innovation in the field.

References

[1] Al-Ayyoub, M., et al. (2018). A system for the automatic recognition of Quran Tajweed rules. *International Journal of Advanced Computer Science and Applications*.

[2] Shaiakhmetov, D., et al. (2025). Evaluation of the Pronunciation of Tajweed Rules Based on DNN as a Step Towards Interactive Recitation Learning. *arXiv preprint arXiv:2503.23470*.

[3] Al-Ghamdi, M. (2005). *Tajweed rules for the holy quran*. Dar Al-Salam.

[4] Ahmad, F. M., et al. (2020). Mispronunciation Detection of Arabic Words using Deep Convolutional Neural Network Features. *Electronics*, 9(6), 963.

[5] Amodei, D., et al. (2016). Deep Speech 2: End-to-End Speech Recognition in English and Mandarin. *International conference on machine learning*.

[6] Ashraf, M. (n.d.). ctc-forced-aligner. *GitHub repository*. Retrieved July 14, 2025, from https://github.com/MahmoudAshraf97/ctc-forced-aligner

[7] Zureiqat, M. (2010). MFCC-based Quranic Verse Recognition System. *2010 Second International Conference on Computational Intelligence, Modelling and Simulation*.

[8] Rahman, A., et al. (2016). Automated Tajweed Checking System for Assisting Children in Quranic Learning. *International Journal on Islamic Applications in Computer Science And Technology*.

[9] Al-Ayyoub, M., et al. (2018). Automatic recognition of quran tajweed rules. *2018 8th International Conference on Computer Science and Information Technology (CSIT)*.

[10] Strik, H., et al. (1997). Acoustic-phonetic features for the automatic detection of mispronunciations. *Journal of the Acoustical Society of America*, 101(5).

A Appendix

This appendix contains supplementary materials.

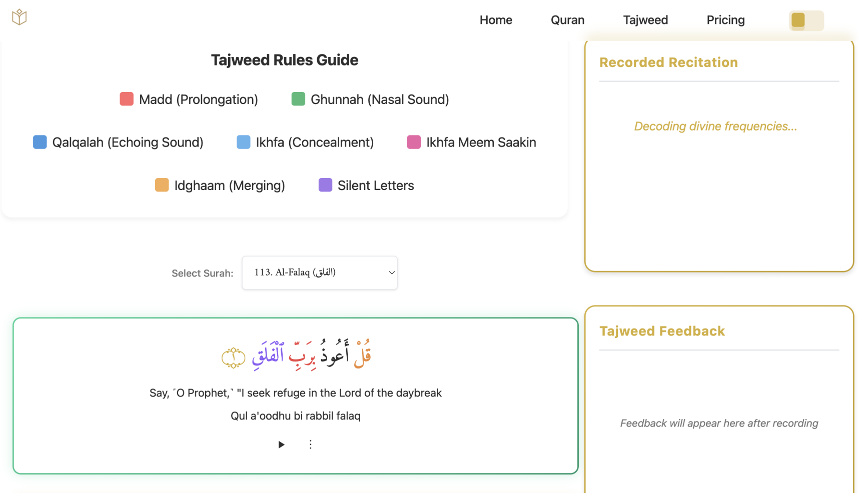

Figure 2: The TajweedAI front-end user interface. The UI displays a guide to Tajweed rules, the selected Surah, and designated areas for recording and feedback.

NeurIPS Paper Checklist

1. **Claims**: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes] The claims regarding the hybrid architecture, iterative methodology, and hard negative mining success are consistently presented and supported by the experimental results in Section 3.

2. **Limitations**: Does the paper discuss the limitations of the work performed by the authors? [Yes] The Abstract and Conclusion explicitly state the model's primary limitation regarding poor generalization to unseen external data (57.14% accuracy) and attribute it to limited data diversity.

3. **Theory assumptions and proofs**: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof? [NA] This paper presents an experimental system and methodology; it does not include theoretical results or proofs.

4. **Experimental result reproducibility**: 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)? [Yes] The paper details the data source (Tarteel QUL), preprocessing tools (NVIDIA ASR model, 'ctc-forced-aligner'), dataset sizes for each experiment, and the hard negative mining methodology.

5. **Open access to data and code**: 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? [No] The paper does not provide public links to the experimental code or the curated/annotated datasets used in the experiments.

6. **Experimental setting/details**: 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? [No] While the paper describes the datasets and models, it omits specific model hyperparameters (e.g., SVM kernel, regularization parameters) and the process for their selection.

7. **Experiment statistical significance**: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments? [No] The experimental results are reported as point estimates of metrics (e.g., accuracy) without error bars, confidence intervals, or statistical significance tests.

8. **Experiments compute resources**: 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? [No] The paper does not provide details on the computational resources (e.g., CPU/GPU type, memory, training time) used for the experiments.

9. **Code of ethics**: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics? [Yes] To the best of our knowledge, the research conforms to all principles outlined in the NeurIPS Code of Ethics.

10. **Broader impacts**: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed? [No] The paper thoroughly discusses the positive societal impact as an educational tool for Quranic recitation but does not address potential negative societal impacts.

11. **Safeguards**: 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? [NA] The paper introduces a specialized educational tool with a low risk for misuse and does not involve the release of a large-scale model or dataset that would require such safeguards.

12. **Licenses for existing assets**: 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? [No] The paper credits the sources of its data (Tarteel QUL) and tools ('ctc-forced-aligner'), but it does not explicitly state the licenses or terms of use for these assets.