Hafs2Vec: A System for the Iqra’Eval Arabic and Qur’anic Phoneme-level Pronunciation Assessment

Ahmed Ibrahim
University of New South Wales
ahmed.ibrahim8165@gmail.com

Abstract

This paper details our submission–Hafs2Vec– to the Iqra’Eval 2025 shared task on Arabic mispronunciation detection. Our system is built upon a wav2vec2-xls-r-1b model, enhanced by two key contributions: a strategic data mixing approach and a custom Qur’anic phonemizer. We augment the official Iqra’Eval training data with 94 hours of professional Qur’anic recitations, creating a balanced dataset that combines learner speech with high-quality acoustic references. To accurately label the reciter data, we developed a custom, Tajweed-aware phonemizer that captures the specific articulation rules of Qur’anic recitation. On the QuranMB test set, our system achieved an F1-score of 46.50% and a high recall of 79.20%.

1 Introduction

The Iqra’Eval 2025 shared task (Kheir et al., 2025) provides a crucial benchmark for advancing Computer-Aided Pronunciation Training in the nuanced domain of Modern Standard Arabic (MSA) and Qur’anic recitation. A primary challenge in developing effective mispronunciation detection systems is the inherent variability in speech data. Learner datasets often contain valuable error patterns but may lack acoustic consistency, while professional recordings offer pristine quality but no examples of common mistakes. Bridging this gap is essential for building models that are both robust and accurate.

To address this challenge, our work introduces two primary contributions. First, we employ a data mixing strategy that combines the 79-hour Iqra’Eval training set with 94 hours of professional Qur’anic recitations. This approach is designed to improve the model’s generalisation by exposing it to a wider range of acoustic conditions, speaking styles, and phonetic details, balancing error diversity with acoustic quality. Second, to enable this strategy, we developed a custom Qur’anic phonemizer (Ibrahim, 2025). This tool generates precise phonetic transcriptions for the professional reciter data by incorporating complex Qur’anic articulation rules governed by Tajweed rules and special symbols within the Qur’anic Uthmani script, which are not accounted for in standard MSA phonemizers, such as Halabi and Wald, 2016.

By integrating these components into a fine-tuned self-supervised learning model framework, our system achieves strong performance. This paper details our methodology, from data preparation and phoneme label generation to model training and evaluation, and provides an in-depth analysis of the system’s performance and error patterns.

2 Methodology

2.1 Data Configuration

We train on a mixture of professional and normal recitations to balance acoustic quality with speaker and error diversity. 28 professional reciters from EveryAyah (Anonymous, 2010) and Qur’anic Universal Library (Tarteel, 2025) are used, filtered to verses of length $\leq$ 10s, yielding $\sim$94 h ($\sim$54k utt.).

The Iqra’Eval training set contributes $\sim$79 h ($\sim$74k utt.) of CommonVoice (Ardila et al., 2020) Arabic speech augmented with Qur’anic recitations.

2.2 Phoneme Label Generation

For the Iqra’Eval data, phoneme labels were provided by the organisers. For the professional reciters set, we generated phoneme labels automatically using a custom Qur’anic phonemizer. This tool takes a verse reference as input and outputs a context-aware, Tajweed-aware phoneme sequence. It expands beyond Modern Standard Arabic phonetics by incorporating Tajweed articulation rules, including Idgham, Iqlab, Ikhfaa, and Qalqala. The phoneme inventory used was matched to the of-

Proceedings of The Third Arabic Natural Language Processing Conference, pages 453–456
November 8-9, 2025 ©2025 Association for Computational Linguistics

Dataset	Utterances	Hours	PER (%)	Sub. (%)	Del. (%)	Ins. (%)
Iqra'Eval Dev (All)	2588	3.4	7.69	4.29	1.62	1.78
Iqra'Eval Dev (Qur'an)	615	–	3.88	1.97	0.84	1.07
Iqra'Eval Dev (MSA)	1973	–	9.50	5.39	1.99	2.11
Professional Reciters	1443	2.6	1.55	0.92	0.37	0.26

Table 1: Phoneme Error Rate (PER) and error type breakdown on development sets.

System	TAR↑	FRR↓	FAR↓	CD↑	Recall↑	Precision↑	F1↑
Organisers' baseline	86.21	13.79	24.44	66.78	75.56	17.67	28.64
Leaderboard Winner (Baic)	**92.09**	**7.91**	34.99	**68.73**	65.01	**37.13**	**47.26**
Our system (Hafs2Vec)	88.40	11.60	**20.80**	62.52	**79.20**	32.92	46.50

Table 2: Mispronunciation detection comparison on the QuranMB test set. TAR: True Acceptance Rate, FRR: False Rejection Rate, FAR: False Acceptance Rate, CD: Correct Diagnosis. Values are percentages. ↓ lower is better, ↑ higher is better.

ficial Iqra'Eval phoneme set — mostly through direct mapping from the phonemizer output, alongside some pre-training and post-training rules.

2.3 Training Configuration

We employed an end-to-end system based on the multilingual facebook/wav2vec2-xls-r-1b (Babu et al., 2021), fine-tuned for 15 epochs with an effective batch size of 352 (22 training batch size × 4 gradient accumulation steps × 4 GPUs). Optimization was performed using AdamW with a learning rate of 3e-5 and a warm-up ratio of 0.1. Experiments were conducted on the University of New South Wales Katana high-performance computing cluster with mixed precision. For inference, greedy decoding was used.

3 Results and Analysis

3.1 Development Set Performance

We evaluated the systems on the Iqra'Eval development set, further categorised into Qur'an and MSA only versions, and the professional reciters development set, consisting of 3 unique reciters and unseen training verses. Table 1 summarises development sets, their PER values and error breakdowns.

On the Iqra'Eval development set, the system achieved a PER of 7.69%, with substitutions (4.29%) as the dominant error type, followed by insertions (1.78%) and deletions (1.62%). Performance is notably better on Qur'anic speech (3.88% PER) than on MSA speech (9.50% PER). This is likely due to the significant variation in style between the CommonVoice MSA data and augmented Qur'anic data. That being said, the MSA subset has approximately 3 times the utterances of Qur'anic subset, so its PER is statistically more stable.

The professional reciters development set shows the lowest PER at 1.55%, reflecting the clarity, consistent articulation, and style match to the training data.

3.2 Test Set Performance

Table 2 compares test set performance of our system with 1. the organisers' baseline system (Kheir et al., 2025) using mHuBERT trained on the CMV-Ar data and 2. the leaderboard-winning system.

Our model achieves a high recall (79.20%) and a low false acceptance rate (20.80%), with a competitive F1-score (46.50%), indicating strong coverage of actual mispronunciations while avoiding many incorrect error detections. The winner leads in F1-score (47.26%) and precision (37.13%), reflecting a more conservative error detection strategy that trades some recall for higher precision.

These results highlight a key trade-off: our system favours high recall and balanced acceptance/rejection behaviour, making it suitable for learner-feedback scenarios where missing genuine errors is more costly than flagging occasional false positives. In contrast, the winning system's higher precision may be advantageous in applications prioritising concise, accurate feedback over exhaustive detection.

3.3 Impact of Data Augmentation

Combining the Iqra'Eval data with professional reciter data introduces greater voice diversity, variation in recitation speed (slow to fast), and exposure to a broader range of acoustic conditions, allowing

Category	Errors	Phonemes	Category PER (%)	Overall PER Contribution (%)
Consonant phonemes	2156	44540	4.84	2.57
Vowel phonemes	4044	36491	11.08	4.82
Shaddah phonemes	238	2818	8.45	0.28
Total	**6438**	**83849**	**–**	**7.69**

Table 3: Error breakdown by phoneme category on the Iqra'Eval development set.

the model to generalise to different recitation styles. Quranic recitation follows distinct modes—such as *mujawwad*, *murattal*, and *hadr*—each with its own tempo and melodic features. Unlike general MDD, we argue that for Quranic MDD it is essential that models are robust to these stylistic differences and that PER is evaluated on multiple test sets representing different styles. A significant decrease in one domain's PER could be more valuable to the overall system at the expense of a slight increase in a different domain's PER, which motivates our use of multiple development sets.

Furthermore, the system demonstrated strong ability to differentiate between the two data domains, successfully avoiding false detection of Tajweed phonemes in the test set, where the Qur'an was recited without Tajweed. Specifically, across the 1,643-utterance test set, only 7 Ikhfaa phonemes and 14 Qalqala phonemes were present, with zero false insertions for all other Tajweed rules. This suggests that the model effectively learned to condition its predictions on the presence or absence of Tajweed articulation, rather than overfitting to Tajweed-rich training data.

Further experimentation could explore the optimal mixing strategy between the two domains, for example by adjusting the ratio of learner to reciter data, applying curriculum learning, or selective sampling.

3.4 Model Selection Findings

Internal experimentation showed that the wav2vec2-xls-r-1b model outperformed the 300M-parameter variant by ∼1.3% absolute PER as well as other models of similar size from the HuBERT (Hsu et al., 2021) and WavLM (Chen et al., 2022) families. This outcome is consistent with the expectation that larger parameter counts enable better representation of complex acoustic patterns, such as those found in Arabic and Qur'anic phonemes. It is reasonable to expect that the 2B-parameter xls-r model could yield further improvements; however, the primary limitations are the increased computational requirements for training and the slower inference speed, alongside risks of overfitting to the training data.

3.5 Categorical Error Analysis

Table 3 presents a breakdown of phoneme errors by broad phoneme categories of consonants, vowels and shaddah phonemes on the development set, allowing for a more robust evaluation of the system (Loweimi et al., 2023). In addition to the total error counts (*Errors*) and the total number of reference phonemes in each category (*Phonemes*), the table reports the *Category PER*, computed for that category alone, and the *Overall PER Contribution*, which reflects the contribution of that category to the overall PER of the dataset.

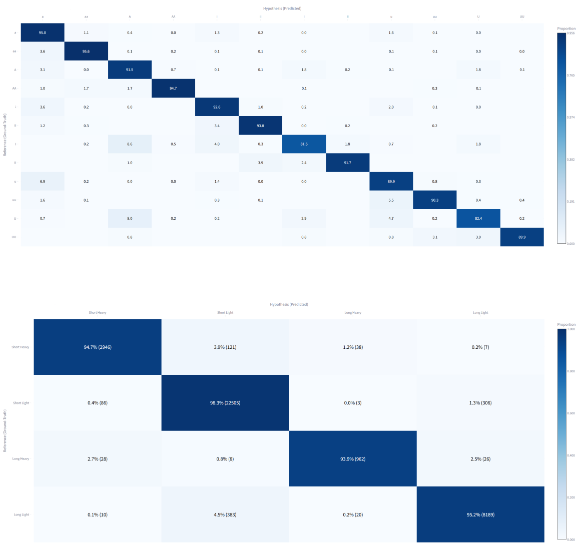

Figure 1: Confusion matrices for vowel phonemes on the development set. (Top) All 12 vowel labels. (Bottom) Vowels grouped by phonetic category: short light = {a u i}, short heavy = {A U I}, long light = {aa uu ii}, long heavy = {AA UU II}.

Vowel phonemes dominate errors, with a PER contribution of 4.82 out of the overall 7.69 (62.8% of all misrecognitions), and exhibiting the highest category PER at 11.08. As seen in the confusion matrices in Figure 1, many of the 12 vowel labels represent acoustically similar sounds (long/short and light/heavy variants), which likely increases confusion, even for human listeners. A possible mitigation would be to reduce the vowel inventory to 6 or 8 broader categories, merging acoustically similar variants while preserving distinctions essential for Arabic speech.

Shaddah (gemination) phonemes, although responsible for only 0.28 PER, have a relatively high category PER (8.45) given their scarcity in the training data: all ten least frequent phonemes in the training data are shaddah forms, each with fewer than 600 occurrences, and the rarest ("EE" and "HH") occur 90 and 94 times respectively. Addressing this severe imbalance may require targeted techniques such as oversampling utterances containing rare phonemes, enforcing balanced phoneme distributions in training batches, or adjusting the loss function to penalise errors on under-represented classes more heavily.

Consonant phonemes are more numerous overall but have a lower category PER (4.84), contributing 33.5% of total errors. These results highlight vowels as the dominant source of phoneme errors, followed by consonants, while rare shaddah phonemes remain a disproportionate challenge given their scarcity in the training data.

4 Conclusion

Our contribution to the Iqra'Eval 2025 shared task demonstrates the effectiveness of a mixed-data training strategy for Arabic mispronunciation detection. By combining learner data and professional, Tajweed-rich recitations with a custom Qur'anic phonemizer, our wav2vec2-xls-r-1b based system is able to generalise across different styles. Our system achieved an F1-score of 46.50%, notable for its high recall (79.20%) on the test set, demonstrating a strong ability to identify genuine errors while minimising incorrect flags.

Our categorical error analysis revealed that acoustically similar vowel phonemes are the primary source of errors ($\sim$63% of all misrecognitions), suggesting that future work focusing on targeted data augmentation or refined vowel inventories could yield significant improvements in the robustness of Arabic pronunciation assessment systems. Furthermore, systematic exploration of optimal data mixing techniques and curriculum learning strategies could further enhance model performance.

5 Acknowledgements

We thank the Iqra'Eval shared task organisers for providing the datasets, evaluation platform, and baseline systems that made this work possible. This research includes computations using the computational cluster Katana supported by Research Technology Services at UNSW Sydney (DOI: 10.26190/669X-A286).

References

Anonymous. 2010. Everyayah dataset. https://everyayah.com/. Online.

Rosana Ardila, Megan Branson, Kelly Davis, Michael Henretty, Michael Kohler, Josh Meyer, Reuben Morais, Lindsay Saunders, Francis M. Tyers, and Gregor Weber. 2020. Common voice: A massively-multilingual speech corpus. *Preprint*, arXiv:1912.06670.

Arun Babu, Changhan Wang, Andros Tjandra, Kushal Lakhotia, Qiantong Xu, Naman Goyal, Kritika Singh, Patrick von Platen, Yatharth Saraf, Juan Pino, Alexei Baevski, Alexis Conneau, and Michael Auli. 2021. XLS-R: self-supervised cross-lingual speech representation learning at scale. *CoRR*, abs/2111.09296.

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, Jian Wu, Long Zhou, Shuo Ren, Yanmin Qian, Yao Qian, Jian Wu, Michael Zeng, Xiangzhan Yu, and Furu Wei. 2022. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1505–1518.

Nawar Halabi and Mike Wald. 2016. Phonetic inventory for an Arabic speech corpus. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 734–738, Portorož, Slovenia. European Language Resources Association (ELRA).

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *Preprint*, arXiv:2106.07447.

Ahmed Ibrahim. 2025. Quranic phonemizer. https://github.com/Hetchy/Quranic-Phonemizer.

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. 2025. Towards a unified benchmark for arabic pronunciation assessment: Quranic recitation as case study. *Preprint*, arXiv:2506.07722.

Erfan Loweimi, Andrea Carmantini, Peter Bell, Steve Renals, and Zoran Cvetkovic. 2023. Phonetic error analysis beyond phone error rate. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:3346–3361.

Tarteel. 2025. Quranic universal library (qul) — recitations and segments data. https://qul.tarteel.ai/resources/recitation.