Iqra'Eval: A Shared Task on Qur'anic Pronunciation Assessment

Yassine El Kheir
DFKI

Amit Meghanani
University of Sheffield

Hawau Olamide Toyin
MBZUAI

Nada Almarwani
Taibah University

Omnia Ibrahim
Alexandria University

Youssef Elshahawy
HUMAIN

Mostafa Shahin
University of New South Wales

Ahmed Ali
HUMAIN

Abstract

We present the findings of the first shared task on Qur'anic pronunciation assessment, which focuses on addressing the unique challenges of evaluating precise pronunciation of Qur'anic recitation. To fill an existing research gap, the **Iqra'Eval 2025 shared task** introduces the first open benchmark for Mispronunciation Detection and Diagnosis (MDD) in Qur'anic recitation, using Modern Standard Arabic (MSA) reading of Qur'anic texts as its case study. The task provides a comprehensive evaluation framework with increasingly complex subtasks: error localization and detailed error diagnosis. Leveraging the recently developed QuranMB benchmark dataset along with auxiliary training resources, this shared task aims to stimulate research in an area of both linguistic and cultural significance while addressing computational challenges in pronunciation assessment.

1 Introduction

The field of Computer-Aided Pronunciation Training (CAPT) and its core component, Mispronunciation Detection and Diagnosis (MDD), have become indispensable tools for self-directed language learners globally (Neri et al., 2008; Rogerson-Revell, 2021). CAPT systems have two main usages: (*i*) pronunciation assessment, where the system is concerned with the errors in the speech segment; (*ii*) pronunciation teaching, where the system is concerned with correcting and guiding the learner to fix mistakes in their pronunciation (Kheir et al., 2023a). Arabic presents unique challenges for CAPT due to its linguistic complexity and diverse varieties. The Arabic phonological system comprises 34 phonemes, including 28 consonants and 6 vowels with distinct short and long forms, which already surpasses the complexity of many Indo-European languages. A particularly salient challenge is posed by complex phonetic structures not commonly found in other languages, such as uvular and pharyngeal consonants, and the subtle but semantically crucial distinction between emphatic and non-emphatic consonants (e.g., / t/ vs. /T/ or /s/ vs. /S/). A slight mispronunciation, such as a substitution between these pairs, can alter the meaning of a word entirely (Kheir et al., 2023b, 2024; Alrashoudi et al., 2025). These challenges in Arabic are **amplified in the domain of Qur'anic recitation**. The recitation of the Holy Qur'an is governed by a strict set of rules known as Tajweed, which dictates the precise articulation of every phoneme, including specific rules for elongation, nasalization (Idgham, Ikhfaa, Iqlab), and bouncing sounds (Qalqala). These rules introduce a layer of phonetic complexity that is absent in Modern Standard Arabic (MSA) and requires specialized models and datasets that can capture these fine-grained acoustic details (Ahmad et al., 2018; Alagrami and Eljazzar, 2020; Rahman et al., 2021; Alsahafi and Asad, 2024). The IqraEval 2025 challenge is motivated by the Unified Benchmark for Arabic Pronunciation Assessment, with Qur'anic recitation as its case study (El Kheir et al., 2025). Building on this foundation, IqraEval 2025 introduces a standardized benchmark supported by carefully curated datasets to tackle the challenges of Arabic MDD. To fill existing gaps, we present the first open benchmark for mispronunciation detection in MSA, specifically focusing on Qur'anic recitation. Our main contributions are:

- **Task Description:** Quranic Mispronunciation Detection and Diagnosis System.
- **Phoneme Set Description:** Detailed phoneme inventory for MSA-based recitation.

Proceedings of The Third Arabic Natural Language Processing Conference, pages 443–452
November 8-9, 2025 ©2025 Association for Computational Linguistics

- **Dataset Release:** Over 80 hours of training and development speech data.
- **Evaluation Framework:** Clearly defined criteria for benchmarking performance.
- **Leaderboard:** The first public leaderboard for Qur'anic Mispronunciation Detection.

2 Iqra'Eval 2025

2.1 Task Description

The Iqra'Eval 2025 shared task focuses on mispronunciation detection and diagnosis in Qur'anic recitation. Given a speech segment and its corresponding reference transcript, the objective is to automatically identify pronunciation errors and localize their positions. In this first iteration of the shared task, the task is framed as a phoneme recognition problem, where the systems are expected to accurately predict the pronounced phonemes in a given MSA-style read Qur'anic Arabic speech recording.

2.2 Dataset and Evaluation

2.2.1 Training Dataset

CV-Ar Dataset This dataset incorporates an 82.37 hours subset of the Common Voice Dataset (Ardila et al., 2019) version 12.0 specifically for MSA Arabic speech recognition. The data set consists of read speech samples collected from a diverse pool of speakers with a well-balanced gender distribution. Fully vowelized versions of the transcriptions developed by El Kheir et al. were used in this shard task. Additionally, the corpus has been augmented with samples drawn from Qur'anic recitations (Alrashoudi et al., 2025).

TTS Augmentation Dataset Introduced in El Kheir et al. to address the scarcity of mispronounced annotated speech data, based on the generative approach techniques demonstrated in Korzekwa et al. 2022. The authors used seven in-house single-speaker TTS systems (5 male and 2 female voices) trained on fully vowelized transcriptions to generate 26 hours of error-free speech (canonical pronunciations) and 26 hours of speech with systematically introduced mispronunciations. The mispronunciation patterns were created by systematically modifying the input of canonical transcripts based on a predefined confusion-pairs matrix derived from phoneme similarity data extracted from Kheir et al. 2022.

2.2.2 Testing Dataset

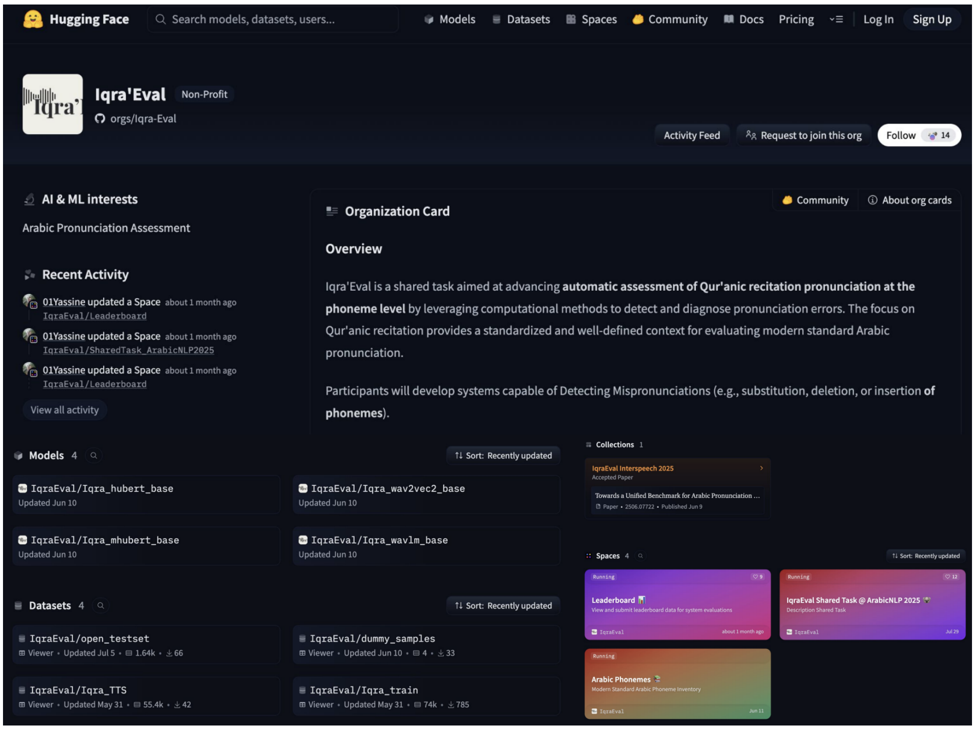
QuranMB The test set introduced by El Kheir et al. consists of 98 verses from the Qur'an, recited by 18 native Arabic speakers (14 females, 4 males), resulting in approximately 2.2 hours of recorded speech. The speakers were instructed to read the text while deliberately producing specified pronunciation errors, which were systematically selected to emulate the most prevalent mispronunciations reported in the literature on common errors in Qur'anic recitation. A custom recording tool was developed to highlight modified text and display additional instructions specifying the type of error to ensure consistency in error production (Alrashoudi et al., 2025). The test set was further annotated by 3 Arabic linguistic annotators.

2.3 Evaluation

We utilize **the specialized phoneme set for Qur'anic Arabic** developed in El Kheir et al., 2025, which builds on the phonetizer introduced by Halabi and Wald, 2016. This set of phonemes includes 62 unique phonemes that account for all MSA sounds, including gemination (the doubling of consonant sounds). The phonemizer has been optimized for phonetic coverage in speech synthesis, employing a greedy algorithm to minimize corpus size while maintaining comprehensive phonetic and prosodic coverage.

Evaluation Metrics Our evaluation protocol adopts the hierarchical structure established in prior mispronunciation detection research (Li et al., 2016; Leung et al., 2019; Kheir et al., 2023a). This framework jointly considers (*i*) the annotated verbatim sequence, (*ii*) the canonical text-dependent reference sequence, and (*iii*) the model prediction. Based on the alignment of these three sources, predictions are categorized into four primary classes:

- **True Accept (TA):** correctly accepted phones that are both annotated and predicted as correct pronunciations.
- **True Reject (TR):** correctly rejected phones that are both annotated and predicted as mispronunciations. These are further exploited to distinguish between *Correct Diagnosis (CD)* and *Error Diagnosis (ED)* depending



Iqra'Eval

orgs/Iqra-Eval

AI & ML interests

Arabic Pronunciation Assessment

Recent Activity

01Yassine updated a Space about 1 month ago
IqraEval/Leaderboard

01Yassine updated a Space about 1 month ago
IqraEval/SharedTask_ArabicNLP2025

01Yassine updated a Space about 1 month ago
IqraEval/Leaderboard

Organization Card

Overview

Iqra'Eval is a shared task aimed at advancing **automatic assessment of Qur'anic recitation pronunciation at the phoneme level** by leveraging computational methods to detect and diagnose pronunciation errors. The focus on Qur'anic recitation provides a standardized and well-defined context for evaluating modern standard Arabic pronunciation.

Participants will develop systems capable of Detecting Mispronunciations (e.g., substitution, deletion, or insertion **of phonemes**).

Models 4

IqraEval/Iqra_hubert_base
Updated Jun 10

IqraEval/Iqra_wav2vec2_base
Updated Jun 10

IqraEval/Iqra_mhubert_base
Updated Jun 10

IqraEval/Iqra_wavlm_base
Updated Jun 10

Datasets 4

IqraEval/open_testset
Viewer • Updated Jul 5 • 1.64k • 66

IqraEval/dummy_samples
Viewer • Updated Jun 10 • 4 • 33

IqraEval/Iqra_TTS
Viewer • Updated May 31 • 55.4k • 42

IqraEval/Iqra_train
Viewer • Updated May 31 • 74k • 785

Collections 1

IqraEval Interspeech 2025
Accepted Paper

Towards a Unified Benchmark for Arabic Pronunciation ...
Paper • 2506.07722 • Published Jun 9

Spaces 4

Leaderboard
View and submit leaderboard data for system evaluations
IqraEval about 1 month ago ♡ 9

IqraEval Shared Task @ ArabicNLP 2025
Description Shared Task
IqraEval Jul 29 ♡ 12

Arabic Phonemes
Modern Standard Arabic Phoneme Inventory
IqraEval Jun 11

Figure 1: Iqra'Eval Shared Task main page.

on whether the predicted phone matches the canonical pronunciation.

- **False Reject (FR):** phones that are actually correct but are incorrectly predicted as mispronunciations.

- **False Accept (FA):** phones that are actually mispronounced but misclassified as correct.

From these four categories, we derive the following error rates.

$$FRR = \frac{FR}{TA + FR} \tag{1}$$

$$FAR = \frac{FA}{FA + TR} \tag{2}$$

$$DER = \frac{ED}{CD + ED} \tag{3}$$

In addition to error rates, we adopt standard diagnostic metrics to evaluate system performance. Precision and Recall are defined as:

$$\text{Precision} = \frac{TR}{TR + FR} \tag{4}$$

$$\text{Recall} = \frac{TR}{TR + FA} = 1 - FAR \tag{5}$$

Finally, the overall performance is summarized using the F1-score, i.e., the harmonic mean of Precision and Recall:

$$F_1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{6}$$

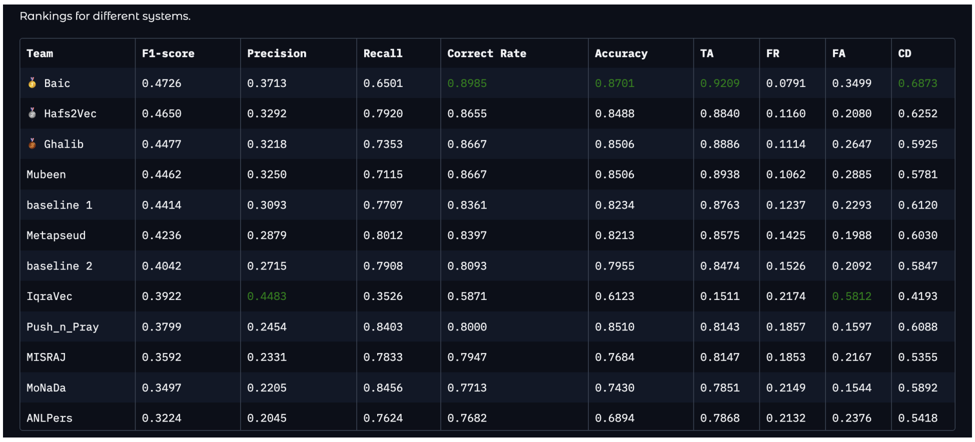
3 Shared Task Teams

Submission Rules All resources for IqraEval are consolidated on the dedicated Hugging Face organization page[1] (see Fig. 1). This page serves as the central hub for datasets, baseline models, reference resources, and evaluation tools. Its main components are summarized as follows:

- **Baseline Models.** Four pretrained SSL models are released for participants: `Iqra_hubert_base`, `Iqra_wav2vec2_base`, `Iqra_mhubert_base`, and `Iqra_wavlm_base`. These provide standardized starting points and ensure comparability across submissions.

- **Datasets.** The page hosts multiple datasets covering training, evaluation, and auxiliary resources:

[1] https://huggingface.co/IqraEval



Rankings for different systems.

Team	F1-score	Precision	Recall	Correct Rate	Accuracy	TA	FR	FA	CD
🥇 Baic	0.4726	0.3713	0.6501	0.8985	0.8701	0.9209	0.0791	0.3499	0.6873
🥈 Hafs2Vec	0.4650	0.3292	0.7920	0.8655	0.8488	0.8840	0.1160	0.2080	0.6252
🥉 Ghalib	0.4477	0.3218	0.7353	0.8667	0.8506	0.8886	0.1114	0.2647	0.5925
Mubeen	0.4462	0.3250	0.7115	0.8667	0.8506	0.8938	0.1062	0.2885	0.5781
baseline 1	0.4414	0.3093	0.7707	0.8361	0.8234	0.8763	0.1237	0.2293	0.6120
Metapseud	0.4236	0.2879	0.8012	0.8397	0.8213	0.8575	0.1425	0.1988	0.6030
baseline 2	0.4042	0.2715	0.7908	0.8093	0.7955	0.8474	0.1526	0.2092	0.5847
IqraVec	0.3922	0.4483	0.3526	0.5871	0.6123	0.1511	0.2174	0.5812	0.4193
Push_n_Pray	0.3799	0.2454	0.8403	0.8000	0.8510	0.8143	0.1857	0.1597	0.6088
MISRAJ	0.3592	0.2331	0.7833	0.7947	0.7684	0.8147	0.1853	0.2167	0.5355
MoNaDa	0.3497	0.2205	0.8456	0.7713	0.7430	0.7851	0.2149	0.1544	0.5892
ANLPers	0.3224	0.2045	0.7624	0.7682	0.6894	0.7868	0.2132	0.2376	0.5418

Figure 2: Iqra'Eval Shared Task Leaderboard.

  - `IqraEval/Iqra_train`: training corpus for system development.
  - `IqraEval/open_testset`: public evaluation split for leaderboard submissions.
  - `IqraEval/Iqra_TTS`: synthetic speech dataset for data augmentation and robustness testing.
  - `IqraEval/dummy_samples`: lightweight set for debugging and format verification.

- **Arabic Phonemes.** A dedicated Space provides an interactive inventory of MSA phonemes, including examples of canonical pronunciations, which supports error diagnosis.

- **Papers.** A collection highlights accepted publications, including the IqraEval Interspeech 2025 paper (El Kheir et al., 2025), which formally describes the benchmark.

- **Leaderboard**[2]**.** An interactive Hugging Face Space is maintained to visualize and compare system outputs. Submitted predictions are automatically evaluated and the leaderboard is updated with human-in-the-loop.

- **Code Samples and Evaluation Scripts.** The organization provides baseline code, sample commands, and the official implementation of evaluation metrics to standardize experimental pipelines and ensure reproducibility.

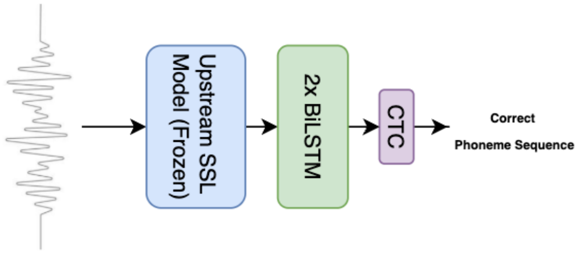


Figure 3: Mispronunciation Detection Modeling Pipeline

- **Submission Workflow.** Participants submit their system outputs in the prescribed CSV format by email following the format provided by the organizers. All valid runs are evaluated automatically, and the results are published on the leaderboard.

Participating Teams A total of 29 teams registered for the shared task. Out of these, 11 teams actively participated in the testing phase and had their systems ranked on the official leaderboard. Among them, 6 teams submitted a system description paper, and 5 of these were accepted for publication in the proceedings of the Iqra'Eval shared task. The participation spanned multiple regions across the globe, including teams from the Middle East, North Africa, Sub-Saharan Africa, South Asia, Europe, North America and Oceania, reflecting the international interest and diversity of the research community engaged in this task.

Baselines We establish two baselines for the Iqra'Eval shared task, as shown in our benchmark (El Kheir et al., 2025), both of which leverage

[2] `https://huggingface.co/spaces/IqraEval/Leaderboard`

Team	Affiliation	Paper Published
ANLPers	Prince Sultan University, Saudi Arabia	✓
BAIC	Applied Innovation Center , Egypt	✓
Greentech	Greentech Apps Foundation, United Kingdom/Bangladesh	
Hafs2Vec	The University of New South Wales, Australia	✓
IqraVec	Imperial College London, United Kingdom	
Metapseud	Independent, Sudan	✓
Misraj Tech	Misraj Technology, Saudi Arabia	
MONADA	- , Tunisia	✓
Mubeen	- , -	
Ghalib	- , -	
Push_n_Pray	Euromed University of Fes, Morocco	

Table 1: List of teams that participated in Iqra'Eval Shared Task.

SSL speech models combined with temporal modeling, as illustrated in Figure 3. Following the SUPERB setup (wen Yang et al., 2021), the SSL encoder parameters are frozen and the layer-wise representations are aggregated through a weighted sum across transformer layers. The resulting features are passed to a model head consisting of a 2-layer, 1024-unit Bi-LSTM trained with Connectionist Temporal Classification (CTC) (Graves et al., 2006) loss on phoneme sequences. During inference, phoneme sequences are obtained using greedy CTC decoding.

- **Baseline 1 (mHuBERT)**: This system employs the multilingual HuBERT (mHuBERT) (Zanon Boito et al., 2024), pretrained on 90,430 hours of speech covering 147 languages. It represents a strong multilingual SSL model suitable for cross-lingual phoneme recognition.

- **Baseline 2 (WavLM).**: This system is based on WavLM (Chen et al., 2022), a 94M-parameter model pretrained on English speech. It provides a monolingual reference point against the multilingual variant.

Our baselines allow us to contrast the effectiveness of multilingual versus monolingual SSL representations for MDD. Below, we provide a brief description for each team system.

3.1 ANLPers:

ANLPers'System (Qandos et al., 2025) is based on `Whisper-large-v3` (Radford et al., 2023), the largest Whisper model with 1.55B parameters and multilingual capabilities. Audio input is resampled to 16 kHz as required by Whisper. The tokenizer is extended with 68 phoneme tokens from (Halabi and Wald, 2016), and the embedding layer is resized accordingly.

The dataset is preprocessed to retain only audio and phoneme attributes. Audio features are extracted using the Whisper feature extractor, and phoneme sequences are encoded as labels. Training is performed using the Hugging Face `transformers` library with a batch size of 4, gradient accumulation of 4 steps, a learning rate of 1×10^{-5}, and 2 epochs.

3.2 BAIC:

BAIC'System (Mattar et al., 2025) is based on Wav2Vec2-BERT (Chung et al., 2021). It employs task-adaptive continued pretraining on large Arabic speech datasets, using phoneme-level labels automatically generated via the Iqra'Eval phonetizer, followed by fine-tuning on the official training data augmented with synthetic Quran recitations created using XTTS-v2. This strategy allows the model to internalize fine-grained phonetic distinctions relevant to mispronunciation detection.

3.3 Hafs2Vec:

This system (Ibrahim, 2025) was trained on two datasets: EveryAyah/QUL, consisting of 94 hours of Quran recitations from 28 professional reciters (filtered to verses under 10 seconds, 54k clips), and the IqraEval training set (79 hours, 74k clips). Phoneme labels for the reciters were generated using a custom Quranic phonemizer that outputs context- and Tajweed-aware phoneme sequences aligned with the IqraEval phoneme set. The model is based on facebook/wav2vec2-xls-r-1b and fine-tuned for 15 epochs with an effective batch size of 352, a learning rate of 3×10^{-5}, AdamW optimization, and CTC loss over the phoneme vocabulary, trained on the UNSW Katana HPC with mixed precision.

3.4 Metapseud:

The submission (Mansour, 2025) applies domain adaptation with multi-stage fine-tuning for phoneme-level Qur'anic mispronunciation detection using Wav2Vec2.0. In the first stage, the pretrained wav2vec2-large-xlsr-53-arabic model is fine-tuned on a large Qur'anic phoneme-annotated dataset (245k recitations), producing a general-purpose phoneme recognizer. In the second stage, the model is further fine-tuned on the official IqraEval training set (79h) to specialize in Qur'anic phoneme structures. Decoding is performed with CTC and beam search, which improves performance on the IqraEval open test set.

3.5 MONADA:

Team MONADA (DAOUD and MESSAOUD, 2025) designed a lightweight system to balance performance with memory efficiency by placing a shallow transformer on top of a pretrained Wav2Vec2.0 feature extractor. Raw audio is processed with the S3PRL Wav2Vec2.0 Base featurizer, producing 768-dimensional frame-level representations, which are then projected into a smaller hidden dimension and fed into a 3-layer transformer encoder with 4 attention heads per layer and a feed-forward size of 1024. The model is trained using CTC loss. Training is conducted for 15 epochs with Adam optimizer (learning rate 3×10^{-4}, cosine annealing scheduler, minimum learning rate 1.5×10^{-5}), dropout of 0.15, and gradient clipping. The best model is selected based on correct rate performance on the development set.

3.6 Mubeen:

The system is based on fine-tuning a Whisper-medium model using the IqraEval training and TTS augmentation data. Only the decoder layers were trained, while the encoder was frozen due to limited hardware and time, using a learning rate of 1×10^{-5} for 2 epochs. An additional fine-tuning pass applied a conservative SpecAugment (Park et al., 2019) strategy, and the resulting models were combined by weight averaging. Inference employed three model configurations with a pairwise WER voting strategy.

3.7 Usubmitted Papers

Out of the 11 participating teams in the IqraEval 2025 Shared Task, 5 teams submitted their test set results to the leaderboard but did not provide any system description or accompanying paper. While their performance contributed to the overall competition rankings, the lack of documentation prevents a detailed analysis of their approaches, training strategies, or architectural choices.

4 Shared Task Results

The overall results for the shared task are in Table 2. Team BAIC (Mattar et al., 2025) presented the best approach, with the best score in 5 of 9 metrics reported. Their model is followed closely by Hafs2Vec (Ibrahim, 2025) and Ghalib. The top 2 approaches included additional training data with BAIC using synthetic data and Hafs2Vec using Quranic recitation from human speakers. The Quranic recitation supplementation data (94 hours) might have affected the model's performance since the training set (79 hours) for Iqra'Eval 2025 is read in MSA style.

5 Lessons from the First Quran Pronunciation Challenge

The submissions revealed three main sources of innovation: model design, data strategies, and training/inference practices. These reflect the community's attempt to balance performance, computational cost, and linguistic specificity.

5.1 Model Innovations

Most teams built on large pretrained encoders such as Whisper, Wav2Vec2, XLS-R, or Wav2Vec2-BERT, demonstrating the effectiveness of transfer learning for Qur'anic mispronunciation detection. Some groups explored lightweight designs, for example MONADA and ShallowTransformer, which placed shallow transformer layers on top of frozen representations to reduce computational cost. Other innovations included extending model vocabularies, such as ANLPers, which

Team	F1-score↑	Precision↑	Recall↑	Correct Rate↑	Accuracy↑	TA	FR	FA	CD
Baic	**0.4726**	0.3713	0.6501	**0.8985**	**0.8701**	**0.9209**	0.0791	0.3499	**0.6873**
Hafs2Vec	0.4650	0.3292	**0.7920**	0.8655	0.8488	0.8840	0.1160	0.2080	0.6252
Ghalib	0.4477	0.3218	0.7353	0.8667	0.8506	0.8886	0.1114	0.2647	0.5925
Mubeen	0.4462	0.3250	0.7115	0.8667	0.8506	0.8938	0.1062	0.2885	0.5781
baseline 1	0.4414	0.3093	0.7707	0.8361	0.8234	0.8763	0.1237	0.2293	0.6120
Metapseud	0.4236	0.2879	0.8012	0.8397	0.8213	0.8575	0.1425	0.1988	0.6030
baseline 2	0.4042	0.2715	0.7908	0.8093	0.7955	0.8474	0.1526	0.2092	0.5847
IqraVec	0.3922	**0.4483**	0.3526	0.5871	0.6123	0.1511	0.2174	**0.5812**	0.4193
Push_n_Pray	0.3799	0.2454	0.8403	0.8000	0.8510	0.8143	0.1857	0.1597	0.6088
MISRAJ	0.3592	0.2331	0.7833	0.7947	0.7684	0.8147	0.1853	0.2167	0.5355
MoNaDa	0.3497	0.2205	0.8456	0.7713	0.7430	0.7851	0.2149	0.1544	0.5892
ANLPers	0.3224	0.2045	0.7624	0.7682	0.6894	0.7868	0.2132	0.2376	0.5418
GreenTech	0.1997	0.1128	**0.8682**	0.5033	0.4585	0.5093	**0.4907**	0.1318	0.4719

Table 2: Evaluation results of different submissions across multiple metrics. Best scores per column are highlighted in bold. Dashed line separates the top 3 submissions.

augmented Whisper's tokenizer with 68 Quran-specific phoneme tokens and resized embeddings accordingly.

5.2 Data Innovations

Several submissions showed that carefully designed resources were central to performance. Hafs2Vec introduced a Tajweed-aware phonemizer to capture recitation rules such as Idgham and Ikhfaa, ensuring that phoneme sequences reflected Qur'anic articulation. BAIC and Mubeen demonstrated the value of TTS-based augmentation using XTTS-v2 to generate synthetic recitations. Hafs2Vec also mixed data from EveryAyah/QUL with the official IqraEval training set to expand speaker and style diversity. In addition, BAIC applied large-scale automatic phoneme labeling with the IqraEval phonetizer to enable task-adaptive continued pretraining on Arabic speech corpora.

5.3 Training and Inference Innovations

Beyond data and model design, training practices had a notable impact. Metapseud applied multi-stage fine-tuning, first adapting to a large Qur'anic phoneme corpus and then specializing on IqraEval. Mubeen selectively fine-tuned only the Whisper decoder layers due to hardware constraints, showing a practical path for parameter-efficient adaptation. Other strategies included the use of SpecAugment and conservative regularization, weight averaging, WER-based voting, and beam search decoding. BAIC highlighted the benefits of task-adaptive pretraining, further reinforcing the importance of domain-specific adaptation.

The summary of innovation by each team is described in the Table 3.

5.4 Emerging patterns

A number of common themes emerged across systems. All teams relied on pretrained SSL encoders, either Whisper or Wav2Vec2 variants, underlining their versatility as general-purpose feature extractors. Quran-specific resources, such as Tajweed-aware phonemizers and synthetic recitations, consistently boosted accuracy and provided linguistic grounding. Training strategies such as multi-stage fine-tuning, selective adaptation, and ensembles yielded measurable gains even without major architectural changes. Finally, the leaderboard revealed different trade-offs in recall versus precision, with some systems favoring high recall for error detection and others prioritizing precision for stricter evaluation.

5.5 Iqra'Eval 2025 Limitations

Limited Linguistic Scope The generalizability of our findings is constrained by the limited linguistic scope of the test data. Although the written Quranic text is standardized, modern spoken Arabic exhibits significant dialectal variation. However, for this shared task, test data was collected exclusively from speakers of the Saudi Arabic dialect. This might limit the generalizability of the results, as the model may fail to capture the rich diversity of the Arabic spoken language.

Team/System	Model Innovation	Data Innovation	Training/Inference Innovation
ANLPers	Whisper-large-v3 with extended phoneme tokenizer	–	HF fine-tuning with resized embeddings
BAIC	Wav2Vec2-BERT backbone	TTS augmentation (XTTS-v2); automatic phoneme labeling	Task-adaptive pretraining; fine-tuning on augmented data
Hafs2Vec	Wav2Vec2-XLS-R-1B	Custom Tajweed-aware phonemizer; mixing EveryAyah/QUL + IqraEval	Large-batch CTC training with AdamW; mixed precision
Metapseud	Wav2Vec2.0 (xlsr-53-arabic)	Large Qur'anic phoneme corpus (245k)	Multi-stage fine-tuning; CTC with beam search
MONADA	Lightweight shallow transformer on Wav2Vec2 features	–	Efficient training with cosine annealing, dropout, clipping
Mubeen	Whisper-medium with frozen encoder; decoder-only training	TTS augmentation	SpecAugment; weight averaging; WER-based voting

Table 3: Summary of innovations from teams that participated in the first Iqra'Eval Shared Task, grouped by model, data, and training methods.

Targets-Specific Common Errors During data collection, speakers were instructed to produce specific pronunciation mistakes deliberately. While this covers some common mistakes, it may not accurately represent the subtle, context-dependent errors that occur in natural recitation, which could limit the model's ability to detect errors in real-world scenarios.

Children IqraEval To the best of our knowledge, there are no publicly available corpora dedicated to children's Qur'an pronunciation learning and recitation assessment. This lack of resources highlights a significant research gap, mainly due to the difficulties of collecting and annotating children's recitation data.

6 Future Work:

We propose the following three directions for future research on our challenge, informed by insights from Iqra'25:

6.1 Task Modelling

Looking ahead, three areas appear particularly promising: (*i*) the balance between precision and recall remains an open challenge: systems must avoid over-flagging errors while still catching subtle mispronunciations; (*ii*) resource creation is essential, especially for rare phonemes and Tajweed-specific contexts where current datasets are imbalanced; (*iii*) efficient adaptation methods such as parameter-efficient fine-tuning, streaming-friendly architectures, or lightweight ensembles could make these models more practical for deployment in real learning settings.

6.2 Data Collection

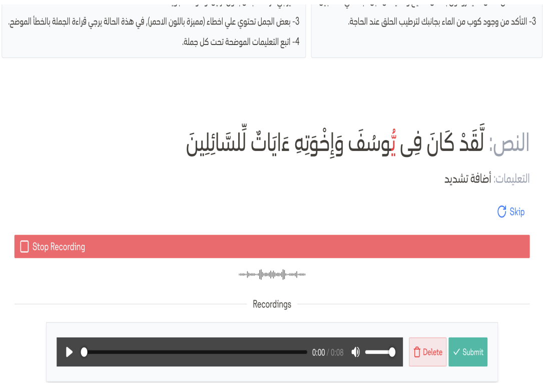
We need more effort to collect and incorporate data from a wide range of Arabic dialects, including but not limited to Egyptian, Levantine, and North African. To capture real-world errors, the next step is to collect recitation audio from a large, diverse group of non-professional reciters, which will then be annotated to identify spontaneous mispronunciations. Further more, we need to develop and release a dedicated corpus for children's Qur'an pronunciation learning and recitation assessment, addressing the current absence of such resources.

6.3 Crowdsourcing platform

We addressed the lack of available data by developing a custom crowd-sourcing platform [3]. This web application allows users to register and provide basic demographic information, including their spoken language, gender, and age. For each

[3] https://quran-data-collection.sanad.ink

sentence, a specific instruction guides the user to introduce a targeted mispronunciation as shown in Figure 4. In cases where a sentence is particularly challenging, no mispronunciation instruction is given, and the user simply reads the sentence as it is. Finally, after a user submits their recordings, we collect the audio data along with their demographic metadata. This information is then prepared for release as a dataset. The collected data will be shared on the Hugging Face platform as part of a shared task, making it accessible to the wider research community. We invite researchers to use our platform to participate and build diveristy corpora for the next Iqra' challenge.



3- التأكد من وجود كوب من الماء بجانبك لترطيب الحلق عند الحاجة.

3- بعض الجمل تحتوي علي اخطاء (مميزة باللون الاحمر), في هذة الحالة يرجي قراءة الجملة بالخطأ الموضح.

4- اتبع التعليمات الموضحة تحت كل جملة.

النص: لَّقَدْ كَانَ فِى يُّوسُفَ وَإِخْوَتِهِ ءَايَاتٌ لِّلسَّائِلِينَ

التعليمات: أضافة تشديد

Figure 4: Data collection screenshot

Acknowledgments

We would like to thank all the annotators who spent hours and hours reviewing our data, and we are grateful to the volunteers who have made lots of recordings across multiple dialects. We thank the Saudi Data & AI Authority (SDAIA) for hosting the Winter School[4] where this work started.

References

Fadzil Ahmad, Saiful Zaimy Yahya, Zuraidi Saad, and Abdul Rahim Ahmad. 2018. Tajweed classification using artificial neural network. In *2018 International Conference on Smart Communications and Networking (SmartNets)*, pages 1–4. IEEE.

Ali M Alagrami and Maged M Eljazzar. 2020. Smartajweed automatic recognition of arabic quranic recitation rules. *arXiv preprint arXiv:2101.04200*.

Norah Alrashoudi, Hend Al-Khalifa, and Yousef Alotaibi. 2025. Improving mispronunciation detection and diagnosis for non-native learners of the arabic language. *Discover Computing*, 28(1):1.

Yousef S Alsahafi and Muhammad Asad. 2024. Empirical study on mispronunciation detection for tajweed rules during quran recitation. In *2024 6th International Conference on Computing and Informatics (ICCI)*, pages 39–45. IEEE.

Rosana Ardila, Megan Branson, Kelly Davis, Michael Henretty, Michael Kohler, Josh Meyer, Reuben Morais, Lindsay Saunders, Francis M Tyers, and Gregor Weber. 2019. Common voice: A massively-multilingual speech corpus. *arXiv preprint arXiv:1912.06670*.

S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao, J. Wu, L. Zhou, S. Ren, Y. Qian, Y. Qian, J. Wu, M. Zeng, X. Yu, and F. 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.

Yu-An Chung, Yu Zhang, Wei Han, Chung-Cheng Chiu, James Qin, Ruoming Pang, and Yonghui Wu. 2021. W2v-bert: Combining contrastive learning and masked language modeling for self-supervised speech pre-training. In *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 244–250. IEEE.

Mohamed Nadhir DAOUD and Mohamed Anouar BEN MESSAOUD. 2025. Phoneme-level mispronunciation detection in quranic recitation using shallowtransformer. In *The Third Arabic Natural Language Processing Conference (ArabicNLP 2025)*, Suzhou. Association for Computational Linguistics.

Yassine El Kheir, Omnia Ibrahim, Amit Meghanani, Nada Almarwani, Hawau 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: Qur'anic Recitation as Case Study. In *Interspeech 2025*, pages 2410–2414.

Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. 2006. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *ICML*, ICML '06, page 369–376, NY, USA. Association for Computing Machinery.

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.

[4] https://www.aiwinterschool.com/

Ahmed Ibrahim. 2025. Hafs2vec: A system for the iqraeval arabic and qur'anic phoneme-level pronunciation assessment shared task. In *The Third Arabic Natural Language Processing Conference (ArabicNLP 2025)*, Suzhou. Association for Computational Linguistics.

Yassine El Kheir, Ahmed Ali, and Shammur Absar Chowdhury. 2023a. Automatic pronunciation assessment–a review. *arXiv preprint arXiv:2310.13974*.

Yassine El Kheir, Shammur Absar Chowdhury, Ahmed Ali, Hamdy Mubarak, and Shazia Afzal. 2022. Speechblender: Speech augmentation framework for mispronunciation data generation. *arXiv preprint arXiv:2211.00923*.

Yassine El Kheir, Fouad Khnaisser, Shammur Absar Chowdhury, Hamdy Mubarak, Shazia Afzal, and Ahmed Ali. 2023b. Qvoice: Arabic speech pronunciation learning application. *arXiv preprint arXiv:2305.07445*.

Yassine El Kheir, Hamdy Mubarak, Ahmed Ali, and Shammur Absar Chowdhury. 2024. Beyond orthography: Automatic recovery of short vowels and dialectal sounds in arabic. *arXiv preprint arXiv:2408.02430*.

Daniel Korzekwa, Jaime Lorenzo-Trueba, Thomas Drugman, and Bozena Kostek. 2022. Computer-assisted pronunciation training—speech synthesis is almost all you need. *Speech Communication*, 142:22–33.

Wai-Kim Leung, Xunying Liu, and Helen Meng. 2019. Cnn-rnn-ctc based end-to-end mispronunciation detection and diagnosis. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8132–8136. IEEE.

Kun Li, Xiaojun Qian, and Helen Meng. 2016. Mispronunciation detection and diagnosis in l2 english speech using multidistribution deep neural networks. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 25(1):193–207.

Ayman Mansour. 2025. Metapseud at iqra'eval: Domain adaptation with multi-stage fine-tuning for phoneme-level qur'anic mispronunciation detection. In *The Third Arabic Natural Language Processing Conference (ArabicNLP 2025)*, Suzhou. Association for Computational Linguistics.

Bassam Mattar, Mohamed Fayed, and Ayman Khalafallah. 2025. Aras2p: Arabic speech-to-phonemes system. In *The Third Arabic Natural Language Processing Conference (ArabicNLP 2025)*, Suzhou. Association for Computational Linguistics.

Ambra Neri, Ornella Mich, Matteo Gerosa, and Diego Giuliani. 2008. The effectiveness of computer assisted pronunciation training for foreign language learning by children. *Computer Assisted Language Learning*, 21(5):393–408.

Daniel S Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D Cubuk, and Quoc V Le. 2019. Specaugment: A simple data augmentation method for automatic speech recognition. *arXiv preprint arXiv:1904.08779*.

Nour Qandos, Serry Sibaee, Samar Ahmad, OMER NACAR, Adel Ammar, Wadii Boulila, and Yasser Alhabashi. 2025. Anplers at iqraeval shared task: Adapting whisper-large-v3 as speech-to-phoneme for qur'anic recitation mispronunciation detection. In *The Third Arabic Natural Language Processing Conference (ArabicNLP 2025)*, Suzhou. Association for Computational Linguistics.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518. PMLR.

Munirah Ab Rahman, Izatul Anis Azwa Kassim, Tasiransurini Ab Rahman, and Siti Zarina Mohd Muji. 2021. Development of automated tajweed checking system for children in learning quran. *Evolution in Electrical and Electronic Engineering*, 2(1):165–176.

Pamela M Rogerson-Revell. 2021. Computer-assisted pronunciation training (capt): Current issues and future directions. *Relc Journal*, 52(1):189–205.

Shu wen Yang, Po-Han Chi, Yung-Sung Chuang, Cheng-I Jeff Lai, and Kushal Lakhotia et al. 2021. Superb: Speech processing universal performance benchmark. In *Interspeech 2021*, pages 1194–1198.

Marcely Zanon Boito, Vivek Iyer, Nikolaos Lagos, Laurent Besacier, and Ioan Calapodescu. 2024. mhubert-147: A compact multilingual hubert model. In *Interspeech 2024*, pages 3939–3943.