Metapseud at Iqra'Eval: Domain Adaptation with Multi-Stage Fine-Tuning for Phoneme-Level Qur'anic Mispronunciation Detection

Ayman Mansour
Independent Researcher
aymanmnsor777@gmail.com

Abstract

This paper presents the Metapseud system designed for the Iqra'Eval shared task, which addresses the automatic assessment of Qur'anic recitation pronunciation, as part of ARABIC-NLP 2025. This system applies multi-stage fine-tuning of Wav2Vec2.0 with curriculum-inspired training, followed by domain adaptation to Qur'anic phoneme annotations. The decoding is improved using beam search with a CTC-based decoder. The results show that staged adaptation achieved a phoneme error rate (PER) of 21% in the development set, and beam search improves the accuracy in the open test set from 76.9% to 82.1%. The findings of this work emphasize the significance of curriculum learning, domain adaptation, and decoding strategies in recognizing mispronunciation in Qur'anic recitation.

1 Introduction

Mispronunciation Detection and Diagnosis (MDD) forms the core of modern Computer-Aided Pronunciation Training (CAPT) systems, providing real-time identification and analysis of learner pronunciation errors. By combining automated speech recognition (ASR), phonetics-driven error detection, and adaptive feedback mechanisms, MDD enables CAPT systems to not only assess pronunciation accuracy but also deliver targeted, pedagogically informed corrective guidance (Neri et al., 2008).

Qur'anic Arabic Automatic Speech Recognition (ASR) presents unique challenges due to its rich phonetic variation, complex Tajweed rules, and significant differences from Modern Standard Arabic (MSA) or dialectal Arabic—placing it in a distinct category often regarded as Classical Arabic (CA) (Habash, 2010). The accurate recitation of the Holy Quran is of profound importance to Muslims worldwide, as it must adhere to precise pronunciation rules (Tajweed), where even minor deviations can alter the meaning entirely. This necessity has motivated initiatives such as the Iqra'Eval shared task (El Kheir et al., 2025), which challenges researchers to develop automatic phoneme-level recognition systems for Qur'anic recitation.

The Metapseud system, which leverages self-supervised ASR model Wav2vec2.0 and combines (1) multistage curriculum-inspired fine-tuning, (2) domain adaptation, and (3) beam search decoding. This strategy is motivated by previous work on curriculum learning and domain adaptation for ASR.

2 Methodology

The methodology focuses primarily on applying Multi-stage fine-tuning by strategically leveraging the self-supervised learning capabilities of Wav2Vec 2.0 (Baevski et al., 2020) by integrating three core techniques: (1) **multistage curriculum-inspired fine-tuning**, (2) **targeted domain adaptation**, and (3) optimized **beam search decoding**. This integrated strategy is motivated by established principles in curriculum learning (Bengio et al., 2009; Platanios et al., 2019) and domain adaptation for speech (Kunze and et al., 2017; Wang and et al., 2021), applying them systematically to a single MDD pipeline for Qur'anic recitation.

2.1 Model Architecture and Foundation

The foundation of this system is the wav2vec2-large-xlsr-53-arabic (Grosman, 2021), based on wav2vec2-large-xlsr-53 a large model pre-trained in 53 languages (Conneau et al., 2020). This model is fine-tuned in Arabic using the train and validation splits of Common Voice 6.1 and Arabic Speech Corpus (Halabi, 2016), while it merely focuses on Arabic ASR, but it provides a robust starting point that subsequently specialize for the

Proceedings of The Third Arabic Natural Language Processing Conference, pages 475–479

target domain.

2.2 Stage-1: General Domain Fine-Tuning (Curriculum Learning)

This stage first exposes the model to Qur'anic recitations to capture prosody and phoneme distributions. Used curriculum-inspired training by first training on a broader Qur'anic dataset that teaches the model general recitation structure and phoneme patterns, by gradually shifting from a general-purpose ASR model to a phoneme-centric Qur'anic recitation model. `Tarteel-ai-EA-DI` dataset (∼245k) is a large and diverse corpus of Qur'anic recitations from various reciters (qurra'). This dataset prioritizes breadth to capture the full range of recitation styles and phonetic variations. The model is fine-tuned on this dataset using a standard Connectionist Temporal Classification (CTC) loss function with a phoneme-level vocabulary.

2.3 Stage-2: Domain Adaptation Fine-Tuning

This stage represents the final step in the curriculum, transitioning the model from a broad understanding to a specialized one. It directly implements domain adaptation (Kunze and et al., 2017; Wang and et al., 2021). To specialize the model for Qur'anic phoneme structures, was fine-tuned the previous model on the provided dataset `Iqra_train` (79 hours) of MSA speech augmented with Qur'anic recitations of Qur'anic phoneme-annotated recitations. This domain adaptation step aligns the model with Qur'anic-specific phoneme distributions, sharpening the model's focus on the specific acoustic features and phonological rules critical for accurate pronunciation evaluation, reducing PER significantly.

2.4 Beam Search Decoding

Beam search decoding is employed using the `PyCTCDecode`[1] library to generate the final phoneme sequences. This method improves upon greedy decoding (Graves et al., 2006; Hannun, 2017) and was chosen to find a more globally optimal sequence compared to the locally optimal stepwise predictions of greedy decoding. The decoding process was implemented using a standard `BeamSearchDecoderCTC` class, initialized with a phoneme vocabulary (`Iqra_train`) specific to this task.

2.5 Evaluation Metrics

The Model performance was evaluated using the hierarchical framework as (Kheir et al., 2023) which assesses both the detection and diagnosis of pronunciation errors, categorizing each predicted phoneme into one of several classes:

- True Accept (**TA**): A correct phoneme is correctly accepted.
- True Reject (**TR**): A mispronounced phoneme is correctly detected as an error.
- False Accept (**FA**): A mispronounced phoneme is incorrectly accepted (i.e., a missed error).
- False Reject (**FR**): A correct phoneme is incorrectly flagged as an error (i.e., a false alarm).

And Diagnosis-Level Categories, Correct Diagnosis (**CD**), Error Diagnosis (**ED**). From these categories, standard information retrieval metrics: **Precision**, **Recall**, and **F-measure** which derived from diagnostic accuracy, and widely used as the performance measures for mispronunciation detection.

$$Precision = \frac{TR}{TR + FR} \tag{1}$$

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

$$F - measure = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \tag{3}$$

3 Data

3.1 Training and Development Data

3.1.1 Every Ayah Diacritized (EA-DI) dataset

The first stage of curriculum learning approach utilized the Every Ayah Diacritized (EA-DI) Phonemized dataset[2], a large-scale, based on Tarteel-Ai's Every Ayah Diacritized (EA-DI) dataset publicly available corpus It encompasses a wide variety of Qur'anic recitations, covering the entire text of the Qur'an from numerous reciters (qurra'). This diversity is crucial for teaching the model the broad acoustic properties and

[1] https://github.com/kensho-technologies/pyctcdecode

[2] https://huggingface.co/datasets/AymanMansour/tarteel-ai-EA-DI-phonemized-Final

phoneme distributions of the domain. Each sample is a rich, diacritized annotation object containing the following key fields:

- Audio: The raw waveform audio signal.
- Transcription: The original orthographic text of the Qur'anic verse.
- Phoneme: The target phoneme sequence for the utterance, generated using a specialized Arabic phonetizer (Kheir et al., 2025). This sequence serves as the primary learning target for phoneme-based recognition model.

3.1.2 Iqra'Eval dataset

The second stage of this work is trained and evaluated using the Iqra'Eval dataset [3], the provided dataset by the shared task, designed for Qur'anic Automatic Speech Recognition (ASR) and pronunciation evaluation. The dataset was utilized in the following predefined splits:

Training Split: Consists of 79 hours of audio. This partition contains a mixture of Modern Standard Arabic (MSA) speech and Qur'anic recitations, providing a curriculum-inspired foundation of general Arabic phonetics before specializing in the target domain.

Development Split: Comprises 3.4 hours of held-out Qur'anic recitations. This split was used exclusively for hyperparameter tuning, validation, and early stopping, ensuring a fair evaluation of the model's generalization capability. Each sample in the dataset follows key fields:

- Audio: The raw waveform audio signal.
- Sentence: The original orthographic text of the Qur'anic verse.
- Index: A unique identifier for the verse.
- Tashkeel_sentence: The fully diacritized text of the verse.
- Phoneme: The target phoneme sequence for the utterance. This sequence serves as the primary learning target for phoneme-based recognition models.

[3] https://huggingface.co/datasets/IqraEval/Iqra_train

3.2 Testing Data

Final evaluation was the IqraEval Open Test datasetet[4]. This dataset is designed as a blind test set; it contains only audio data without ground truth transcriptions. The dataset consists of $\approx$ 2 h, with deliberate errors and human annotations, Predictions generated on this set are submitted to the Iqra'Eval organizers for evaluation scoring.

4 Results

4.1 Development Results

Stage-1 This curriculum setup helped stabilize training and improved the model's ability to generalize phoneme boundaries. After fine-tuning, the model achieved PER $\approx$ 0.54 on the held-out development data(EA-DI), establishing a strong baseline. **Stage-2** performed domain adaptation by further fine-tuning the stage-1 model on the Iqra_train dataset, which represents the official shared task domain. This stage achieved PER $\approx$ 0.21.

Model	Dataset	PER
Stage-1	EA-DI	0.54
Stage-2	Iqra_train	0.21

Table 1: Models Results on development set.

4.2 Open Test Results

Finally, beam search decoding was applied, yielding further gains at the sequence level. On the open test set, the best submission achieved an F1 score of 0.4236, with an accuracy of 0.8213 .

4.3 Qualitative Results

In this section a qualitative analysis is conducted based on examples from the development set. A comparative analysis of the outputs of both models againstst ground truth (Figure 1) indicates that deletions constitute the primary error type, with a limited number of perfect matches. Additionally, Figure 2 summarizes the character pairs that cause the most confusion.
Performance Statistics for 100 samples:

- Perfect Matches: 11 (11.0%)
- BS Improved: 11 (11.0%)
- BS Same Error: 68 (68.0%)

[4] https://huggingface.co/datasets/IqraEval/open_testset

	F1-score(%)↑	Recall(%)↑	Precision(%)↑	CR(%)↑	Accu(%)↑	TA(%)↑	FR(%)↓	FA(%)↓	CD(%)↑
Baseline 1	**44.14**	**30.93**	77.07	83.61	**82.34**	**87.63**	**12.37**	22.93	**61.2**
Baseline 2	40.42	27.15	79.08	80.93	79.55	84.74	15.26	20.92	58.47
Stage-2	40.74	27.5	78.61	83.28	76.89	85.1	14.9	21.398	59.4
Stage-2 (BS)	42.36	28.79	**80.12**	**83.97**	82.13	85.75	14.25	**19.88**	60.3

Table 2: Experimental Results. ↓ Lower is better, ↑ Higher is better.

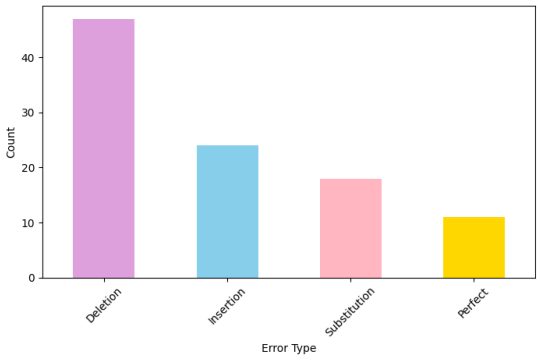


Figure 1: Error Type Distribution

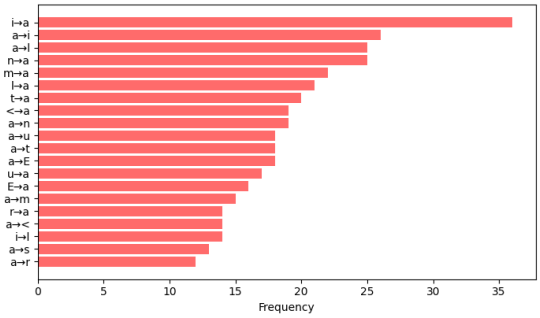


Figure 2: Top 20 Character Confusions (GT→Model)

- BS Worse: 10 (10.0%)

The error analysis reveals that the beam search decoding strategy provided minimal performance gains, often reproducing the same errors as the base model. In addition, a strong positive correlation was observed between the length of an utterance and the number of errors.

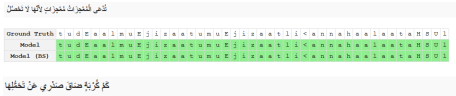
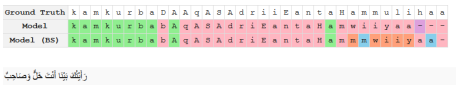
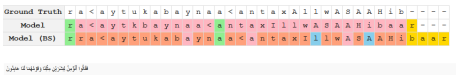
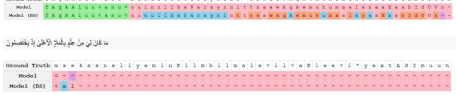
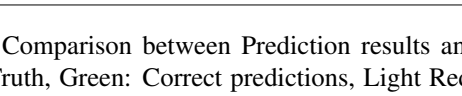
Sample	Results
1	
2	
3	
4	
5	

Table 3: Comparison between Prediction results and Ground Truth, Green: Correct predictions, Light Red: Substitution errors, Blue: Characters corrected by Beam Search, Purple: Deletion errors, Gold: Insertion errors, Light Orange: Different errors in BS vs regular model

5 Discussion

Curriculum Learning. The staged approach validates curriculum-inspired fine-tuning (Bengio et al., 2009), as the general Qur'anic recitation training improved domain-specific adaptation.

Domain Adaptation. Without Stage-1, direct fine-tuning on IqraEval resulted in poor generalization. Adaptation through staged training aligns well with previous findings (Kunze and et al., 2017; Wang and et al., 2021).

Decoding. Beam search mitigated concatenation errors and improved phoneme sequences.

6 Conclusion

Two-stage fine-tuning pipeline was presented with domain adaptation and beam search decoding for Qur'anic ASR. Future work may include tajweed-aware decoding, phoneme-level language models, and adaptive curriculum schedules.

References

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460.

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. 2009. Curriculum learning. In *Proceedings of ICML*.

Alexis Conneau, Alexei Baevski, Ronan Collobert, Abdelrahman Mohamed, and Michael Auli. 2020. Unsupervised cross-lingual representation learning for speech recognition. *arXiv preprint arXiv:2006.13979*.

Yassine El Kheir, Amit Meghanani, Hawau Olamide Toyin, Nada Almarwani, Omnia Ibrahim, Youssef Elshahawy, Mostafa Shahin, and Ahmed Ali. 2025. Iqra'eval: A shared task on qur'anic pronunciation assessment. In *Proceedings of the Third Arabic Natural Language Processing Conference*. Association for Computational Linguistics.

Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. 2006. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd international conference on Machine learning*, pages 369–376.

Jonatas Grosman. 2021. Fine-tuned XLSR-53 large model for speech recognition in Arabic. https://huggingface.co/jonatasgrosman/wav2vec2-large-xlsr-53-arabic.

Nizar Y Habash. 2010. *Introduction to Arabic natural language processing*. Morgan & Claypool Publishers.

Nawar Halabi. 2016. *Modern standard Arabic phonetics for speech synthesis*. Ph.D. thesis, University of Southampton.

Awni Hannun. 2017. Sequence modeling with ctc. *Distill*. Https://distill.pub/2017/ctc.

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

Yassine El Kheir, Omnia Ibrahim, Amit Meghanani, Nada Almarwani, Hawau Olamide Toyin, Sadeen Alharbi, Modar Alfadly, Lamya Alkanhal, Ibrahim Selim, Shehab Elbatal, and 1 others. 2025. Towards a unified benchmark for arabic pronunciation assessment: Quranic recitation as case study. *arXiv preprint arXiv:2506.07722*.

Julius Kunze and et al. 2017. Transfer learning for speech recognition on a budget. In *Interspeech*.

Ambra Neri, Catia Cucchiarini, and Helmer Strik. 2008. The effectiveness of computer-based speech corrective feedback for improving segmental quality in l2 dutch. *ReCALL*, 20(2):225–243.

Emmanouil A Platanios, Otil Stretcu, Graham Neubig, Barnabás Póczos, and Tom M Mitchell. 2019. Competence-based curriculum learning for neural machine translation. In *Proceedings of NAACL*.

Changhan Wang and et al. 2021. Fine-tuning self-supervised speech models with limited data. In *ICASSP*.