ANPLers at IqraEval Shared task: Adapting Whisper-large-v3 as Speech-to-Phoneme for Qur'anic Recitation Mispronunciation Detection

Nour Qandous[1], Serry Sibaee[2*], Samar Ahmed[1], Omer Nacar[3], Yasser Al-Habashi[2]
Adel Ammar[2], Wadii Boulila[2]
[1]NAMAA, Riyadh, Saudi Arabia
[2]Prince Sultan University, Riyadh, Saudi Arabia
[3]Tuwaiq Academy – Tuwaiq Research and Development Center, Riyadh, Saudi Arabia

Abstract

Mispronunciation detection at the phoneme level provides detailed feedback for Quranic reciters. Standard speech-to-text models cannot capture subtle differences in letter pronunciation; thus, developing a speech-to-phoneme system is essential. Prior works have mainly explored encoder-only models. In this work, we adapt Whisper-large-v3 on the IqraEval dataset. Our experimental results show that the proposed system achieved an F1-score of 0.3224, an accuracy of 0.6894, and a high recall of 0.7624. These results highlight promising directions for further research and development in phoneme-level mispronunciation detection.

1 Introduction

Computer-aided language Learning (CALL) employs computer technologies to aid in language acquisition and pronunciation. Accurate pronunciation is critical in the context of Arabic language learning, particularly for Quranic recitation, as it has both linguistic and religious significance.

Speech-to-Phoneme (STP) transcribe audio to tokens of phonemes; it differs from conventional Speech-to-Text (STT) systems that transcribe audio into tokens of words. Unlike STT, STP provides a finer-grained phonetic representation that is essential for precise pronunciation feedback and error detection, particularly relevant for Quranic recitation where subtle phonetic distinctions alter meaning and correctness.

Given the importance of phoneme-level accuracy for Quranic recitation, STP systems offer a promising approach to support learners in mastering Quranic pronunciation by analyzing their recitation audio record, detecting any fine-grained error, then provide a corrective feedback.

In this work, we fine-tuned **Whisper-large-v3** as an STP model to help detect mispronunciation of the Quranic recitation at the phoneme level. This paper is organized as follows: Section 2 reviews related works in mispronunciation detection using phonemes in Quranic recitation. Section 3 describes the proposed system and the experimental setup. Section 4 reports and analyzes the results. Section 5 discusses the results and provides insights. Finally, Section 6 concludes the paper and outlines possible directions for future work.

2 Related Works

Building on prior work in Quranic recitation recognition(Al-Zaro et al., 2025) developed a phoneme-based speech recognition system for Quranic recitation using the DeepSpeech architecture. They proposed a phoneme list of 53 units, covering consonants, vowels, and Tajweed-specific phonemes. The system was trained on a combined dataset consisting of a proprietary corpus from EqraTech and the open-source ASR Tarteel dataset, due to the lack of publicly available phoneme-level datasets for the Quran, totaling approximately 550 hours of audio. Evaluation results reported a phoneme error rate (PER) of 7.4%, a word phoneme error rate (WPER) of 27.97%, and a word error rate (WER) of 3.92% at the `imlā'ī` (spelling) word level.

Similarly(Calik et al., 2023) presented an ensemble-based framework for detecting mispronunciations of Arabic phonemes, particularly in the context of Quranic pronunciation, utilizing machine learning techniques including SVM, k-NN, and decision trees. The system also used feature extraction methods to enhance language learning through computer-assisted tools. It employed both traditional and ensemble learning-based approaches for evaluation. An accuracy of 95.9% was achieved using a voting classifier with mel-spectrogram features.

Extending the focus to Tajweed-specific errors(Harere and Jallad, 2023) developed an au-

Proceedings of The Third Arabic Natural Language Processing Conference, pages 464–468

tomatic mispronunciation detection system for Quranic recitation based on Tajweed rules (Separate Stretching, Tight Noon, and Hide), using the QDAT dataset, a public dataset containing over 1,500 audio samples of correct and incorrect recitations. The system addressed the shortage of qualified human supervisors by leveraging deep learning, specifically Long Short-Term Memory (LSTM) networks. The model achieved high accuracy rates of 96% for Separate Stretching, 95% for Hide, and 96% for Tight Noon. In a broader context of language learning (Algabri et al., 2022) applied deep learning to develop a versatile high-performance assisted pronunciation system for the detection, diagnosis, and generation of articulatory feedback for non-native Arabic learners. It used YOLO-based object detection for phoneme and articulatory feature recognition and employed a CNN-RNN-CTC model to provide feedback. The system achieved a phoneme error rate (PER) of 3.83% in the phoneme recognition task, an F1-score of 70.53% in the mispronunciation detection and diagnosis task, and a detection error rate (DER) of 2.6% in the articulatory feature detection task.

Most recently, (Şükrü Selim Çalık et al., 2024) introduced a novel framework for mispronunciation detection of Arabic phonemes using audio-based transformer models such as Squeezed and Efficient Wav2Vec (SEW), Hidden-Unit BERT (HUBERT), Wav2Vec, and UNI-SPEECH. The study is considered the first to comprehensively explore Arabic phoneme mispronunciations using these models. A dataset consisting of 29 Arabic phonemes, including 8 hafiz sounds, was collected from 11 speakers and supplemented with additional samples from YouTube. The results demonstrated that the UNI-SPEECH model achieved superior performance. Moreover, the proposed framework was designed to be speaker-independent, allowing for general applicability without the need for individual speaker enrollment.

Previous studies show us that they focused on using encoder-only architecture in solving such a problem. Therefore, we are exploring a new direction by investigating encoder–decoder architecture, namely the Whisper model, to detect mispronunciation on phoneme level.

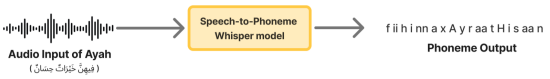

Figure 1: Overview of the proposed speech-to-phoneme system.

3 Methodology

3.1 Dataset

We utilized the complete training and evaluation subsets from the IqraEval dataset (El Kheir et al., 2025) for model training, yielding a total of 73,990 records and approximately 82.4 hours of audio. The dataset consists of short Arabic audio recorded by multiple speakers, paired with the corresponding sentences and detailed phoneme transcriptions. Each sample also includes metadata such as a unique identifier and the sentence with tashkeel. Although the dataset contains multiple attributes, we only utilized the audio and phoneme attributes for the speech-to-phoneme (STP) task.

3.2 Proposed System

Figure 1 presents the proposed speech-to-phoneme (STP) system, in which an audio recording of recitation is transcribed into a phoneme sequence using our STP model. We base our system on **Whisper-large-v3**, the largest Whisper variant with 1550M parameters and multilingual support. The reason for selecting **Whisper-large-v3** is that it achieved outstanding performance on the Arabic Automatic Speech Recognition (ASR) leaderboard (Wang et al., 2024) with an average word error rate of 0.3686, and because it is faster to fine-tune compared to other exceptional models. Our assumption was that if an STT model achieved high performance on processing Arabic audio into words, it would get promising performance in transcribing Arabic Audio into phonemes, and we will discuss this assumption later in the discussion section 5.

To enable phoneme-level prediction, we extend the original Whisper tokenizer by incorporating 68 additional phoneme tokens from (Halabi and Wald, 2016), and resize the embedding layer to match the new vocabulary. This modification yields an adapted Whisper model capable of generating phoneme sequences directly from audio inputs.

Breaking down the system as illustrated in Figure 2, the Whisper feature extractor first computes log-mel spectrograms from the input audio. These

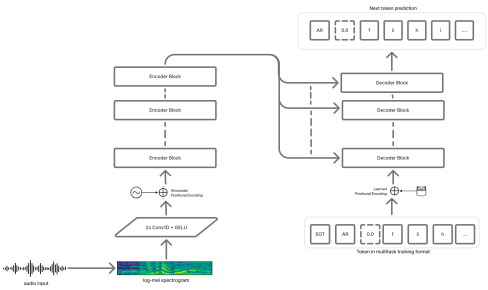

Figure 2: Whisper architecture for the speech-to-phoneme task, using a custom tokenizer with 68 phoneme tokens.

spectrograms are processed by the encoder to produce high-level acoustic representations. The decoder then predicts the output sequence token-by-token, conditioned on the preceding phoneme tokens, until the complete phoneme transcription is generated. During training, phoneme sequences are used as target labels to guide the prediction process.

3.3 Hyperparameter Optimization

This training setup optimizes for efficient fine-tuning by using a small batch size with gradient accumulation to simulate a larger effective batch, a low learning rate for stable updates, and mixed precision (fp16) to reduce memory usage. It saves checkpoints frequently while limiting total saved models, evaluates performance by word error rate (WER) after each epoch, and aims to minimize WER for best model selection. The training was run on 3 A100 80GB GPUs, leveraging a batch size of 4 per device and gradient accumulation over 3 steps to optimize memory usage and training efficiency.

4 Results

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

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

$$F_1 = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall} \tag{3}$$

Table 1 presents the evaluation metrics of our system on the Iqraeval test set. True Rejects (TR) are the percentage of mispronunciations correctly transcribed as phonemes, while False Accepts (FA) represent the percentage of failures to transcribe mispronunciations. True Accepts (TA) demonstrate the correct transcription of correct pronunciation, while False Rejects (FR) are the percentage of incorrect transcription of correct pronunciation. See Equations 1, 2, and 3 for more context.

As shown in Table 1, the model demonstrated a high recall of 0.7624 and a correct rate of 0.7682, indicating strong capability in transcribing mispronounced phonemes and covering phoneme variations effectively. However, the relatively low precision of 0.2045 suggest that the system suffers from a high number of False Rejects (FR), yielding a low F1-score of 0.3224.

The true acceptance rate (TA) of 0.7868 shows that the majority of correctly pronounced phonemes are transcribed correctly, yet the False Accept (FA) of 0.2376 reveals that a significant portion of mispronunciations remains undetected. Accuracy and correct detection rate (CD), at 0.6894 and 0.5418, respectively, indicate moderate overall classification quality.

These results demonstrate that our system can be optimized to transcribe fine-grained features, and there is room to improve the precision and reduce the false rejects to achieve a more balanced and effective phoneme recognition performance.

Metric	Value
F1-score	0.3224
Precision	0.2045
Recall	0.7624
Correct Rate	0.7682
Accuracy	0.6894
TA	0.7868
FR	0.2132
FA	0.2376
CD	0.5418

Table 1: Test set results for the proposed system.

5 Discussion

Table 2 shows model predictions for three recordings of the same Ayah [78-Al Imran] with supposedly different pronunciations. Although the model should generate three different transcriptions for these three recordings, it generates identical transcriptions for the first two and slightly different transcriptions for the third one. As illustrated in bold in Table 2, the difference was in a **single phoneme** in the word **yalwun**. All three recordings

ID	Prediction
00000_00013	w a n m i n h u m l a f a r ii q AA y a l w u n **u** E a l s i n a t a h u m b i l k i t aa b
00000_00013143	w a n m i n h u m l a f a r ii q AA y a l w u n **u** E a l s i n a t a h u m b i l k i t aa b
00000_00013343	w a n m i n h u m l a f a r ii q AA y a l w u n **a** E a l s i n a t a h u m b i l k i t aa b

Table 2: Outputs of the proposed system for three recordings identified by ID metadata from the IqraEval test set, all corresponding to Ayah 78 of Surah Al-Imran.

could appear identical to a regular Arabic listener, except that the end of the word **yalwun** needs to be clearer from the speaker, or, as said in Tajweed, the sound needs more duration. So from our perspective, it is difficult for an Arabic listener to predict the correct phoneme.

Moreover, prior works demonstrated that using CTC-based models with self-supervised encoders such as Wav2Vec 2.0 achieved high performance in phoneme-level mispronunciation detection (Kheir et al., 2025). Our approach adopts an encoder–decoder paradigm with **Whisper-large-v3**. The experimental results show the effectiveness of our approach. The downside is the adaptation limitations that require significant hardware and prevent real-time operation, as the model has 1500M parameters. Quantizing the model or using smaller versions and increasing the number of epochs could provide more robust performance and a more reliable system.

5.1 Analysis of the dataset

Our comprehensive analysis of the Arabic voice dataset revealed several critical quality issues that warrant careful consideration for the shared task implementation. Technical artifacts were prevalent throughout the collection, with numerous audio samples exhibiting signal truncation and cutting issues that compromise the integrity of the speech data. A substantial portion of recordings demonstrated incorrect application of tahreek (diacritical markings) at sentence endpoints, deviating from standard Arabic phonological conventions for proper vocalization. While the dataset predominantly consists of Quranic recitations, we identified instances of mispronunciation that diverge from canonical tajweed principles, potentially introducing phonetic inconsistencies in model training. Temporal irregularities were observed across the corpus, with speaking rates varying significantly from normal conversational pace, which may adversely affect automatic speech recognition performance and temporal alignment algorithms. Furthermore, we detected grammatical errors within the spoken content and critical misalignments between reference transcriptions and their corresponding audio files, representing fundamental data integrity challenges. These systematic quality control issues necessitate robust preprocessing pipelines and filtering mechanisms to ensure dataset reliability and maintain the validity of experimental results in Arabic speech processing applications.

6 Conclusion

Mispronunciation detection is a challenging task, and it becomes even harder in Quranic recitation scenarios; thus, phoneme-level mispronunciation detection is essential. In this paper, we explored the use of the Whisper model and addressed this problem as a speech-to-phoneme task. Our results reveal a substantial gap between a recall of 0.7624 and a precision of 0.2045, indicating that while the model effectively identifies a wide range of mispronounced phonemes, it frequently misclassifies correctly pronounced ones. A limitation of this work results from the hardware requirements that prevented us from experimenting with more epochs. For future work, we encourage researchers to explore encoder–decoder architectures for phoneme-level mispronunciation detection, investigate Nvidia Conformer CTC Arabic models, which combine the Conformer architecture with CTC and have shown high Arabic STT performance (Wang et al., 2024). These directions could further advance research in this area, provide a more reliable system to improve Quranic recitations for Muslims, and inspire new ideas for mispronunciation detection.

References

Samah Al-Zaro, Mahmoud Al-Ayyoub, and Osama Al-Khaleel. 2025. Speaker-independent phoneme-based automatic quranic speech recognition using deep learning. *IEEE Access*, 13:125881–125896.

Mohammed Algabri, Hassan Mathkour, Mansour Alsulaiman, and Mohamed A Bencherif. 2022. Mispronunciation detection and diagnosis with articulatory-

level feedback generation for non-native arabic speech. *Mathematics*, 10(15):2727.

Sukru Selim Calik, Ayhan Kucukmanisa, and Zeynep Hilal Kilimci. 2023. An ensemble-based framework for mispronunciation detection of arabic phonemes. *Preprint*, arXiv:2301.01378.

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.

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).

Ahmad Al Harere and Khloud Al Jallad. 2023. Mispronunciation detection of basic quranic recitation rules using deep learning. *Preprint*, arXiv:2305.06429.

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.

Yingzhi Wang, Anas Alhmoud, and Muhammad Alqurishi. 2024. Open universal arabic asr leaderboard. *arXiv preprint arXiv:2412.13788*.

Şükrü Selim Çalık, Ayhan Küçükmanisa, and Zeynep Hilal Kilimci. 2024. A novel framework for mispronunciation detection of arabic phonemes using audio-oriented transformer models. *Applied Acoustics*, 215:109711.