Automatic Pronunciation Error Detection and Correction of the Holy Quran's Learners Using Deep Learning

Assessing spoken language is challenging, and quantifying pronunciation metrics for machine learning models is even harder. However, for the Holy Quran, this task is simplified by the rigorous recitation rules (tajweed) established by Muslim scholars, enabling highly effective assessment. Despite this advantage, the scarcity of high-quality annotated data remains a significant barrier.

In this work, we bridge these gaps by introducing:

(1) A 98% automated pipeline to produce high-quality Quranic datasets - encompassing: Collection of recitations from expert reciters, Segmentation at pause points (waqf) using our fine-tuned wav2vec2-BERT model, Transcription of segments, Transcript verification via our novel Tasmeea algorithm;

(2) 850+ hours of audio (~300K annotated utterances);

(3) A novel ASR-based approach for pronunciation error detection, utilizing our custom Quran Phonetic Script (QPS) to encode Tajweed rules (unlike the IPA standard for Modern Standard Arabic). QPS uses a two-level script: (Phoneme level): Encodes Arabic letters with short/long vowels. (Sifa level): Encodes articulation characteristics of every phoneme. We further include comprehensive modeling with our novel multi-level CTC Model which achieved 0.16% average Phoneme Error Rate (PER) on the testset. We release all code, data, and models as open-source: https://obadx.github.io/prepare-quran-dataset/