COMPUTER SCIENCE RESEARCH BULLETIN
ISSN 2774-3659 (Online Media)
Vol 5, No 4, June 2025 | Pages 778-787
https://hostjournals.com/bulletincsr
DOI: 10.47065/bulletincsr.v5i4.561

# Evaluating End-to-End ASR for Qur'an Recitation Using Whispers in Low Resource Settings

**Abdullah Azzam*, Ichsan Taufik, Aldy Rialdy Atmadja**

Informatics Department, Faculty of Science and Technology, Universitas Islam Negeri Sunan Gunung Djati, Bandung, Indonesia
Email: [1,*]abdullahazzam2199@gmail.com, [2]ichsan_taufiq@gmail.com, [3]aldyrialdy@uinsgd.ac.id
Correspondence Author Email: abdullahazzam2199@gmail.com

**Abstract**−This study investigated the use of End-to-End Automatic Speech Recognition (E2E ASR) for Qur'an recitation under low resource conditions using the Whisper model. This study follows the CRISP-DM methodology, starting with defining the research gap and preparing a curated dataset of 200 verses from Juz 30. These verses were chosen because of their short and consistent structure, allowing for efficient experimentation. Audio and transcription pairs are verified and cleaned to ensure alignment and quality. The modeling was done using Whisper in Google Colaboratory, leveraging its pre-trained architecture to reduce training time and computing costs. Evaluations use the Character Error Rate (CER) metric to measure transcription accuracy. The results showed that Whisper achieved an average CER of 0.142, corresponding to a transcription accuracy of about 85%. However, the average processing time per father is 11 seconds, almost double the time it takes for a human readout. Although Whisper provides strong accuracy for Arabic transcription, its runtime efficiency remains a challenge in real-time applications. This research contributes reproducible channels, validated datasets, and performance benchmarks for future studies of the Qur'anic ASR under computational constraints.

**Keywords**: End-to-end ASR; Recitation of the Qur'an; Whispering Models; Low-Resource Speech Recognition; Character Error Rate

## 1. INTRODUCTION

In the study of Islamic law, the Qur'an is the main source in determining the law, occupying the first position as the basis of Islamic law. The Qur'an is the main argument in the formation of Islamic law. So that every determination and formation of the basis whether in different places, times and conditions, Muslims will continue to make the Qur'an the first basis.[1]

In other words, the Qur'an is a guide to life for Muslims around the world. Therefore, understanding and practicing its teachings is the obligation of every Muslim. To be able to understand the Qur'an, the first step is to be able to read it[2]. The ability to read the Quran is the first step to be able to learn the Quran.

The above urgency is enough to place the Qur'an as one of the first things to be learned as a Muslim[3]. One way to learn the Qur'an is to memorize the Qur'an. This method is also one of the methods used to maintain the validity of the Qur'an from various changes and defects[4]. This is also in accordance with the statement of QS.15:9 which means "Verily We have sent down the Qur'an and indeed We (also) will guard it".

The process of memorizing the Quran has many problems ranging from memorization methods[5] or the existence of memorization partners to be able to correct the reading of a person who is memorized[6]. The existence of a partner or if viewed in substance is to correct the memorized reading of a memorizer, this is very important because it could be that a memorizer does memorize but what is memorized is wrong because it has never been corrected by the partner or the tool to correct it.

This process requires a long step to be realized as a solution to the above problems. The problem above is how to make a machine that can correct the reading of the Qur'an correctly. Machines cannot directly read sounds coming out of humans, but they can convert sounds into machine-readable data or files, which are usually text files, hence the term End-to-End Automatic Speech Recognition (ASR) [7].

This research focuses on solving one of the problems of the steps mentioned above, namely End-to-end ASR (Automatic Speech Recognition) [7]. End-to-end ASR can be defined as the process of obtaining a transcription (text) of speech or sound using only one model[7]. ASR can be used as an initial stage to correct a person's reading by a computer, before which the computer must be able to read the input from the person's voice, which is where the role of ASR lies. At this stage, ASR has been widely used in some business processes [8] or even in some languages, but the use of ASR for Quranic datasets is still not wide[9].

A deeper and less explored area in Automatic Speech Recognition (ASR) is End-to-End ASR (E2E), which has received relatively limited research attention compared to traditional ASR systems[10]. Unlike conventional ASR which involves multiple modules such as acoustic modeling, pronunciation lexicon, and language modeling, the E2E ASR system directly maps raw speech input to text output in a single integrated neural network architecture[11]. These models have attracted interest due to their architectural simplicity, integrated training objectives, and reduced reliance on handmade features and intermediate representations [12], [13]. However, despite its potential, the E2E ASR model is still understudied, mainly because its development typically requires substantial computing resources and extensive data for training [14]. While large-scale speech datasets have become increasingly available, many remain inadequately pre-processed—containing noise, misalignment, and inconsistencies—which can significantly degrade the model's accuracy and generalization[12], [13].

End-to-End (E2E) Automatic Speech Recognition (ASR) has emerged as a promising alternative to traditional ASR systems. However, research in E2E ASR is still relatively limited due to the complexity of training and the need for large and high-quality datasets, especially using Quranic datasets, although many datasets are available, they often contain a lot of noise, background interference, inconsistent pronunciation, and lack of standard pre-processing, which complicates model training and reduces transcription accuracy[9], [15]. In addition, processing a large corpus of the Qur'an requires considerable computing resources, especially if there is no curated data[16].

In recent years, the development of Automatic Speech Recognition (ASR) has increasingly shifted to an End-to-End (E2E) approach, which eliminates the modular dependencies found in traditional hybrid systems. Prabhavalkar *et al.* [12]and Li [13] note that E2E ASR architectures—such as CTC, RNN-Transducer, and attention-based models—offer simplified pipelines and reduce reliance on handmade components. However, despite its promise, E2E ASR remains underexplored compared to traditional ASR frameworks.

Some works have identified the limitations of current ASR systems when applied in a low-resource or domain-specific context. Alharbi *et al.* [8] emphasizes that most ASR research assumes access to large, clean datasets, which is not always practical. Rahman *et al.* [17] observed that even in languages with substantial corpora, noise and inconsistent pre-processing are common problems, affecting recognition performance. Furthermore, Alsayadi *et al.* [18] argues that the Arabic ASR, especially for Classical Arabic or the Qur'an—faces additional challenges due to phonological complexity and the scarcity of clean, annotated audio-text pairs.

The research gaps identified in this study highlight several key challenges within the field of Automatic Speech Recognition (ASR) that remain underexplored. First, End-to-End ASR (E2E ASR) has received relatively limited attention compared to traditional ASR models, particularly in the context of complex tasks such as Qur'anic recitation. The simplicity and integrated architecture of E2E ASR offer promising potential; however, its application in this domain remains understudied. Second, ASR research typically relies on large datasets, making it computationally intensive and resource-demanding, especially in low-resource environments. This study focuses on addressing this challenge by developing efficient ASR solutions tailored to limited computational resources. Lastly, many existing datasets, even those for languages with substantial corpora, have not been adequately pre-processed, resulting in issues such as noise, misalignment, and inconsistencies that hinder the performance of ASR models.

The research problems stemming from the identified research gaps that this research seeks to address include several key issues. Firstly, the deeper aspect of Automatic Speech Recognition (ASR), specifically End-to-End ASR, has received relatively little research attention, leaving a gap in understanding and development in this area. Secondly, ASR research often depends on large datasets, which makes the research computationally intensive, requiring significant resources to process the data efficiently. Lastly, even for languages with large datasets, many of these datasets have not been properly pre-processed beforehand, which affects the quality and reliability of ASR systems. This study aims to address these gaps by focusing on these challenges. Given the challenges mentioned above, this study seeks to contribute by addressing various issues in the field of end-to-end automatic speech recognition (ASR). In particular, this study aims to explore solutions to the lack of research on end-to-end ASR systems, especially in low-resource settings. This research also focuses on the development of models or frameworks that can achieve high accuracy with limited data availability. Finally, the study highlights the importance of utilizing fully processed datasets to minimize external noise and interference.

Thus, the contribution of this research can be summarized in several key points. First, it provides valuable research on End-to-End Automatic Speech Recognition (ASR), focusing on its application to Qur'anic recitation. Second, the study identifies which models or frameworks are capable of achieving high accuracy in low-resource condition using evaluation metrics. Lastly, the research delivers a pre-processed Qur'an recitation dataset, which ensures the verification and alignment between the transcripts and audio files.

## 2. RESEARCH METHODOLOGY

CRISP-DM (Cross Industry Standard Process for Data Mining) is a widely accepted industry-independent process model that structures the exploration and modeling of data through different iterative phases [19]. Its flexibility and clear procedural steps make it a perfect fit for research projects aimed at developing End-to-End Automatic Speech Recognition (ASR) systems, where systematic data analysis and modeling are essential to achieve high accuracy under low-resource conditions [20]. In this study, CRISP-DM was used by following the stages of Business Understanding, Data Understanding, Data Preparation, Modeling, and Evaluation. The deployment phase was excluded, as the research focused on the development and evaluation of models and the delivery of pre-processed Qur'anic recitation datasets, rather than the operational implementation of a system, which is in line with the finding that academic applications of CRISP-DM often do not require implementation when the primary goal is model optimization rather than production use [21]. The following is the CRISP-DM workflow.

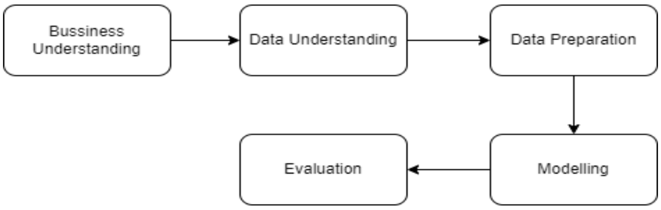

**Figure 1.** CRISP-DM Workflow

### 2.1 Business Understanding

The Business Understanding phase in CRISP-DM involves defining the business, identifying current conditions, and presenting a project plan. This step is considered essential to align technical work with business objectives and serves as a strategic starting point for data science projects[22]. As explained earlier in the paragraph on the Business Understanding stage, the scope of this research is limited to the mentioned research contributions, whereas the current conditions regarding end-to-end automatic speech recognition for the recitation of the Qur'an are outlined in the research gaps. The project plan in this study emphasizes the need for resources for the training process, in particular the need for low resource conditions. This approach minimizes computational demands while aiming to achieve high model accuracy. The core methodology involves utilizing small pre-processed data sets to train models to perform the transcription of Qur'anic speeches into Arabic texts. For model evaluation, the Character Error Rate (CER) will be used as the primary evaluation metric.

### 2.2 Data Understanding

In the Data Understanding phase, it is common to identify and document the data sources to be used, describe the data in a structured manner, verify the quality of the data, and formulate hypotheses based on initial observations [22]. This activity aims to ensure the relevance and reliability of the data set, while also guiding the next steps in data preparation and modeling through proper assumptions about the characteristics of the data.

In this study, datasets were obtained from the following sources: https://www.kaggle.com/datasets/omartariq612/quran-reciters/data. The dataset is organized into MP3 audio files, separated into eight folders that correspond to each reader. Each folder contains a verse-level audio file, where the file name indicates the surah and the verse number—for example, 001001 represents Surah 1, Verse 1, and 002001 represents Surah 2, Verse 1. The accompanying metadata in CSV format includes the number of verses per surah and the Arabic transcription for each verse.

The dataset used in this study consisted of recordings from eight different reciters, with audio quality ranging from 40 kbps to 192 kbps and a total size of about 10-11 GB. Each reading is segmented per verse, with file names arranged numerically to reflect the surah and father (e.g., 001001 show Surah Al-Fatihah, Father 1). Each file varies in duration and size depending on the content and bitrate, as described in Figure 4.

In addition to audio files, the dataset includes multiple CSV transcription files. It contains a variety of forms of Arabic text—some with diacritics, some without, and some with waqf symbols—providing flexibility to choose a transcript format based on experimental needs. The dataset also includes ayah_count.csv files that list the number of verses in each surah. Upon examination, the recordings from the folder Husary_128kbps found to be of very high quality, with clear pronunciation and minimal background noise, as described in Figure 6.

Exploration of the preliminary data shows that the audio quality is generally good, especially for recordings by Hussary reciters, which show clear and understandable pronunciation. However, there are variations in transcription files: some contain full diacritics (harakat), while others are non-vowel. All transcriptions are written in Arabic script. Regardless of systematic naming conventions, further verification may be required to ensure accurate alignment between audio files and appropriate textual transcriptions, especially where naming alone may not guarantee semantic or sequential consistency.

As an initial hypothesis, although the datasets appear to be well-organized, model adjustments may still be needed to accommodate the structural and linguistic characteristics of the data. This study also seeks to examine the performance of the model under low resource conditions, which is specifically defined in terms of data size (file volume). This study will explore whether high transcription accuracy can still be achieved when using smaller datasets, thus addressing efficiency in scenarios with limited storage or processing capabilities.

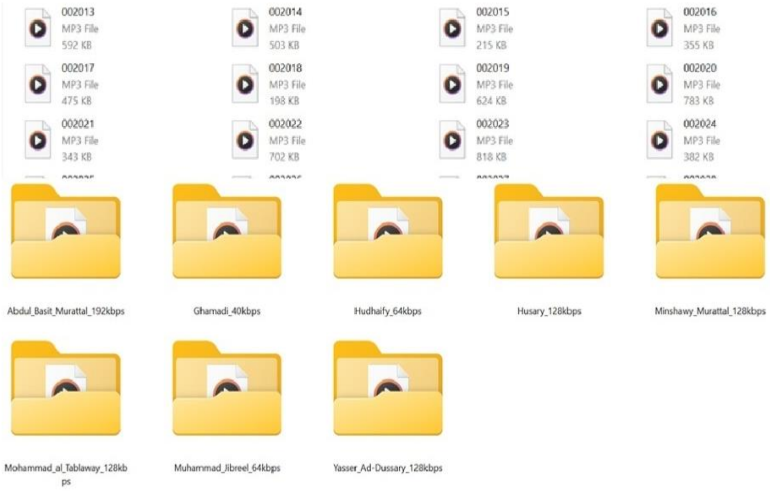

**Figure 2.** Audio Files and Folders

### 2.3 Data Preparation (data pre-processing)

Data preparation is a critical phase that focuses on transforming raw data into a final dataset that can be used directly by the model. This stage typically involves a series of systematic steps, including data selection, cleaning, structuring, integration, and formatting. In addition, it requires a clear definition of data inputs and outputs, as well as the transformations required to align the data with the model's requirements. The ultimate goal of this process is to ensure that the data set is coherent, consistent, and tailored to support optimal model performance[22]. In this study, the data selection process involves the selection of 30 letters from Juz 30 of the Quran. Juz 30 was chosen as a subset for the experiment due to the shorter average verse length, which better supports rapid testing and ASR performance measurement on short speech.

After understanding the structure of the dataset, pre-processing steps are performed to support the effectiveness of the experiment. This includes developing verification scripts to ensure each audio file accurately matches its transcript. Although the data is well organized, manual verification is also performed on a subset to avoid potential misalignments, as shown in Table 1. Only non-diacritical, non-waqf transcript versions were selected for use, in line with the novelty of the study operating in low-resource linguistic conditions, also illustrated in Table 1.

The verification process results in a valid audio transcript pair of the last 30 chapters of Juz 30. These verified pairs are saved into a structured CSV file for further processing. Verification confirms that no entries are missing or mismatched, allowing for a complete and consistent experimental trial consisting of the final 30 chapters of Juz 30. The results of this verification are illustrated in Table 1. With the pre-processing completed and the data set standardized by excluding diacritics (harakat) and tajweed annotations—focusing only on standard Arabic scripts—the experiment proceeded to the modeling phase. The pre-processed datasets are then stored in Google Drive as part of the research contribution of this research.

**Table 1.** Example of a Verified Transcript Dataset without Diacritics

| Surah | Father | Text | Condition |
|---|---|---|---|
| 112 | 2 | Allah Al, Samad | TRUE |
| 112 | 3 | He was not born and was not born | TRUE |
| 112 | 4 | And there is no one like him | TRUE |
| 113 | 1 | Say I take refuge in the God of division | TRUE |
| 113 | 2 | Of what evil He created | TRUE |
| 113 | 3 | And from the wickedness of the wicked if he falls | TRUE |
| 113 | 4 | It was a jet crime under contract | TRUE |
| 113 | 5 | And of the evil of the envious if he envies | TRUE |
| 114 | 1 | .Say I take refuge in the Lord of the universe | TRUE |
| 114 | 2 | The king of the people | TRUE |
| 114 | 3 | God of man | TRUE |

### 2.4 Modeling

The modeling phase consists of several key steps such as selecting appropriate modeling techniques, building the model, assessing its performance, and refining it through training and testing. This phase involves selecting the right tools, setting appropriate parameters, and implementing modeling methods that align with the data and business goals. The main focus of this stage is to develop and validate models that can produce accurate and reliable outputs[22].

#### 2.4.1 Whisper Model

In the modeling phase of this study, Whisper was chosen as the end-to-end automatic speech recognition (ASR) core model. Whisper is a multilingual, multitasking system trained on 680,000 hours of supervised audio-text pairs, enabling robust performance in low-resource settings [23]. Its encoder-decoder architecture supports transcription without the need for language-specific pre-processing steps, making it suitable for structured datasets such as Qur'an readings [24]. While Whispers have been explored in previous ASR tasks of the Qur'an, the consistency of transcription across all verses and the reader remains an open challenge [25]. Therefore, this study evaluated Whisper using a curated and pre-processed Qur'anic audio dataset to assess its effectiveness under low-resource conditions. Compared to previous models such as the VOSK, which resulted in low accuracy on similar tasks, the Whisper offers a more promising basis for further improvement [26]. Based on the code applied, the workflow of this research is arranged as follows. The initial step involves verifying the existence and alignment of the audio file with the appropriate transcript to ensure the consistency of the data set. Once verified, the audio is processed using Whisper. Whisper directly converts digital audio waveforms into text through a unified architecture that includes an encoder and a decoder.

The encoder converts the raw audio into a high-level representation of features, which is then processed by the decoder to produce a transcription. This end-to-end process allows the model to capture complex patterns between speech and text without intermediate linguistic annotations. Whispers are used solely for the recitation of the Qur'an into standard Arabic script. The resulting text is evaluated against the reference transcript using the Character Error Rate (CER), which measures transcription accuracy by calculating insertions, deletions, and substitutions.

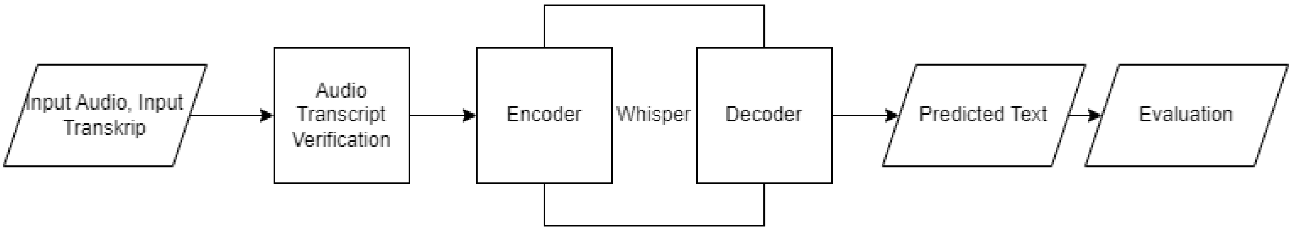

**Figure 3.** Research Experiment Workflow

### 2.4.2 Low Power Supply Conditions

This study addresses the low-resource conditions in End-to-End Automatic Speech Recognition (E2E ASR) by minimizing data volume and computational complexity. An audio portion of the curated Qur'an recitation is used, focusing on high-quality recordings while eliminating linguistic features such as diacritics and tajweed rules, which are known to increase phonetic complexity and computational costs [9].

To adapt to computational constraints, the study used Whisper, a multilingual ASR model, an encoder-decoder trained on 680,000 hours of labeled audio-text pairs. Whisper has shown strong performance in a variety of languages and low-resource domains, including structured religious texts such as the Qur'an [23], [24]. Instead of full retraining, the model is applied through direct inference or light refinement to reduce resource demand.

### 2.4.3 Tools

The study leveraged Google Colaboratory as the primary development environment, leveraging its cloud-based GPUs for efficient ASR processing. Google Drive is used for dataset management, storing Qur'anic audio files, and verified transcripts. Python is the core programming language, with panda handling metadata alignment. OpenAI's Whisper model supports end-to-end Arabic transcription, suitable for low-resource conditions. The jiwer library calculates the Character Error Rate (CER), which is appropriate for the Arabic language due to the complexity of the script. Supporting tools such as os and tqdm facilitate file access and feedback processing. In addition, matplotlib is used to visualize CER trends and transcription performance across sentences. Together, these tools form a lightweight and reproducible channel for Qur'an reading, transcription, and evaluation in resource-constrained settings.

### 2.5 Evaluation

This evaluation phase focuses on a critical assessment of the performance and validity of the model developed prior to implementation. This stage typically includes determining evaluation metrics, reviewing the modeling process, evaluating the results, visualizing the results, and determining next steps[22]. This activity aims to ensure that the model performs well in accordance with standard metrics and aligns with the original research objectives, ultimately guiding decision-making based on the results achieved.

Ultimately, the Character Error Rate (CER) is an evaluation metric in Automatic Speech Recognition (ASR) that measures the difference in character levels between the predicted transcription and its reference. CERs are widely used in multilingual ASR evaluations due to their effectiveness in handling languages that are morphologically complex and that do not have clear word boundaries, such as Arabic, Chinese, and Thai. Compared to Word Error Rate (WER), CER is considered more consistent, especially in cross-language scenarios. The CER is calculated by summing the number of character replacements (S), deletions (D), and insertions (I), then dividing this number by the total number (N) of characters in the reference text. In other words, the CER is equal to (substitution + deletion + insertion) divided by the total number of reference characters. Studies have shown that CER correlates more strongly with human judgment across languages than WER[27]. The metric formula is as shown below.

$$\frac{(S+D+I)}{N} \tag{1}$$

### 2.6 End-to-end ASR

End-to-End Automatic Speech Recognition (E2E ASR) has become a significant advancement in the field of speech processing, offering a unified approach by integrating all components of Automatic Speech Recognition (ASR) into a single neural network. Traditional ASR systems typically rely on multiple models, such as acoustic models, language models, and lexicons, whereas E2E models map raw speech directly to text without these intermediate steps, simplifying the ASR pipeline and improving efficiency. The shift to E2E models has shown remarkable performance improvements, particularly in reducing word error rates (WER) and increasing generalization across various tasks. Studies have shown that E2E models outperform traditional hybrid systems in many benchmarks[12].  highlighting a significant reduction in WER through deep learning-based approaches, which directly optimize the ASR system as a whole.

Despite these advantages, E2E ASR systems face specific challenges, particularly in dealing with languages that have complex phonological structures, such as Arabic. For domain-specific applications like Qur'anic recitation, the task becomes even more complex. Current models such as Connectionist Temporal Classification (CTC), Recurrent Neural

Network Transducer (RNN-T), and Attention-based Encoder-Decoder (AED) have made significant strides but still require large, pre-processed datasets for training. This reliance on large datasets becomes problematic in low-resource settings. Furthermore, E2E models often lack explicit alignment between the input speech and the output text, which is a critical challenge in ASR, especially when training data is scarce or noisy. Attention mechanisms have been introduced to improve this alignment[13] solutions for better handling phonological complexities and noisy datasets are still under investigation. Further advancements are needed to optimize these models for low-resource environments and specific domains like Qur'anic recitation.

### 2.7 Recitation of the Qur'an

The Recitation of the Qur'an (Qira'at) is a central aspect of Islamic practice, essential not only for spiritual guidance but also for preserving the accuracy and integrity of the sacred text. The art of Qur'anic recitation is governed by rules of tajweed and Qira'at, which ensure that the Qur'an is read with the correct pronunciation, rhythm, and intonation. The significance of proper recitation has been emphasized in both historical and contemporary Islamic scholarship, as it is a key method of transmitting the Qur'an's meanings accurately. In the context of technological advancements, the End-to-End Automatic Speech Recognition (E2E ASR) models have been explored for their potential to assist in Qur'anic recitation transcription, offering an innovative solution to address the challenges of pronunciation accuracy and phonetic complexity in Arabic[28] . However, while these models have shown promising results in general ASR tasks, their application to Qur'anic recitation still requires further refinement, particularly in low-resource settings where computational power and data availability are limited.

The Recitation of the Qur'an has also seen its adaptation and reinforcement through various educational initiatives[29]. discuss the importance of Qur'anic recitation education as a tool for character building and spiritual development in Muslim communities, especially in areas like southern Thailand. This method includes structured recitation classes, competitions, and public performances, all aimed at enhancing students' abilities to recite the Qur'an with proper pronunciation. Such efforts align with the growing interest in leveraging E2E ASR systems to support Qur'anic education by offering real-time feedback to students. These technologies can help bridge gaps in learning and offer personalized assistance for Qur'anic memorization and recitation practices, particularly in regions where traditional learning resources may be scarce. Nevertheless, the challenge remains to enhance the performance of ASR models, especially in the case of Qur'anic recitation, where phonological intricacies and pronunciation variances can significantly affect the accuracy of transcription.

### 2.8 Whisper Model

The Whisper Model, developed by OpenAI, represents a significant step forward in the development of Automatic Speech Recognition (ASR) systems. Whisper has been trained on an extensive multilingual dataset, including over 680,000 hours of speech data from various languages, making it highly versatile for transcription, translation, and even real-time speech recognition tasks [30]. This large training dataset allows Whisper to demonstrate improved robustness against various accents, background noise, and technical jargon, which is particularly important in applications like Qur'anic recitation where phonetic and semantic precision is crucial. Whisper’s End-to-End ASR architecture simplifies the typical multi-module setup of traditional ASR systems, offering a more efficient way to process audio into transcriptions by directly mapping speech to text [31].

However, Whisper is not without its limitations. While it excels in transcription accuracy, it has been reported to sometimes generate hallucinated transcriptions, i.e., text that is not present in the original speech. This can significantly affect transcription reliability, which is a critical issue for tasks such as Qur'anic recitation where the accuracy of the transcribed text is non-negotiable. The performance of Whisper also varies based on the dialect and accent of the speaker. Research has shown that Whisper performs better with standard accents like American English compared to more diverse non-native accents, highlighting a gap in its performance for non-native Qur'anic recitation, which often involves speakers with diverse linguistic backgrounds [30]. Additionally, Whisper’s larger models, though more accurate, come with increased latency and require significant computational resources, which may hinder real-time applications like Qur'anic recitation feedback systems [30]. This challenge is particularly relevant when Whisper is deployed in low-resource environments, such as mobile devices or educational settings where computational power may be limited.

### 2.9 Low Resource Speech Recognition

Low-resource speech recognition refers to the development of ASR systems in settings where transcribed audio data is scarce, language resources are limited, and computational infrastructure is constrained. These conditions are common for minority languages, dialects, or domain-specific content such as liturgical or classical texts. Unlike high-resource scenarios where models can be trained on hundreds or thousands of hours of labeled speech, low-resource settings often rely on only a few hours of annotated data. That such limitations significantly hinder the ability of large models like Whisper to generalize effectively unless adapted through targeted fine-tuning strategies[32]. Their work shows that even with under 20 hours of labeled data, meaningful performance gains can be achieved by freezing encoder layers and applying parameter-efficient methods, thus preserving computational feasibility while avoiding overfitting.

Further complicating low-resource ASR is the issue of domain mismatch and language dissimilarity, especially when cross-lingual transfer is employed. The solution address this by proposing the use of donor languages that share acoustic characteristics with the target language to support continued pretraining[33]. Their introduction of the Acoustic Token Distribution Similarity (ATDS) metric enables the selection of donor languages based on empirical acoustic similarities rather than typological assumptions. This method allows for effective adaptation even when the target language lacks sufficient transcribed or even untranscribed corpora. These approaches underscore that successful low-resource ASR depends not only on model architecture but also on intelligent data curation, pretraining adaptation, and efficient parameter utilization—principles that are crucial when developing ASR systems for domains such as Qur'anic recitation.

### 2.10 Character Error Rate

Character Error Rate (CER) has emerged as a crucial evaluation metric in assessing the performance of Automatic Speech Recognition (ASR) systems, particularly when confronted with multilingual contexts and the need for character-level evaluation precision. This metric quantifies the proportion of erroneous characters in a hypothesis transcription relative to the reference transcription, accounting for character insertions, deletions, and substitutions. The advantages of CER become more pronounced in languages with complex morphology or those lacking clear word boundaries, such as Arabic. Some explicitly advocate for the use of CER as the primary metric in multilingual ASR evaluation[27]. They argue that CER successfully avoids various challenges faced by Word Error Rate (WER) and demonstrates greater consistency across diverse writing systems, which is highly relevant for multilingual contexts where WER can be biased or less informative.

Nevertheless, accurate CER measurement heavily relies on the availability of reliable ground truth labels. There is solution to address the challenges in evaluating ASR models, especially when labeled data from diverse domains and testing conditions is limited, which hinders the assessment of models' true generalization capabilities[34]. As a solution, they propose a novel label-free approach to approximate ASR performance metrics, including CER. This approach leverages multimodal embeddings in a unified space for speech and transcription representations, combined with a high-quality proxy model. This indicates that while CER is a robust metric, ongoing research strives to develop more efficient and dependable evaluation methods, even in scenarios where ground truth might not be fully available or easily accessible.

## 3. RESULTS AND DISCUSSION

A total of 327 verses of the last 30 surahs of Juz 30 were evaluated. The average processing time per father is 9.02 seconds. Individual father's processing time varied slightly but remained within a narrow range, indicating stable computing performance across the dataset. With a total processing time of 327 verses, which is 2950 or approximately 50 minutes. This is explained as in Figure 4.

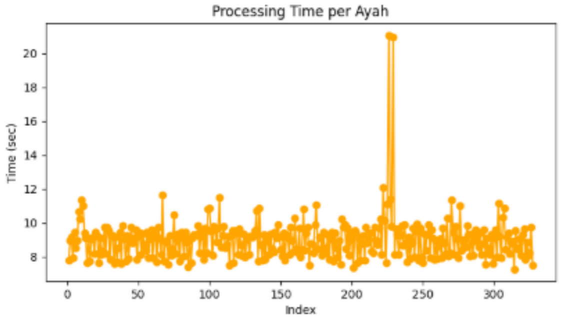

**Figure 4.** Process Time per Dad

The average RAM usage per father was recorded at 0.0023 MB. The value of RAM usage was consistent across all samples, with no significant outliers or memory spikes observed during processing. As shown in figure 5.

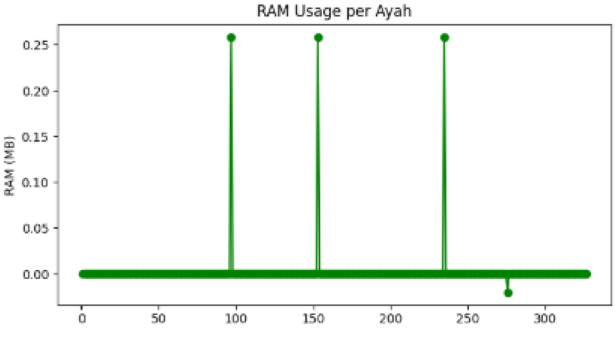

**Figure 5.** Average RAM Usage

The average Character Error Rate (CER) across fathers is 0.144. Most fathers achieved a CER of 0.0, while a small subset showed a CER value instead of zero, contributing to the overall average. This distribution shows that most predictions match the basic truth text exactly, with some deviations in certain cases. As shown in figure 6

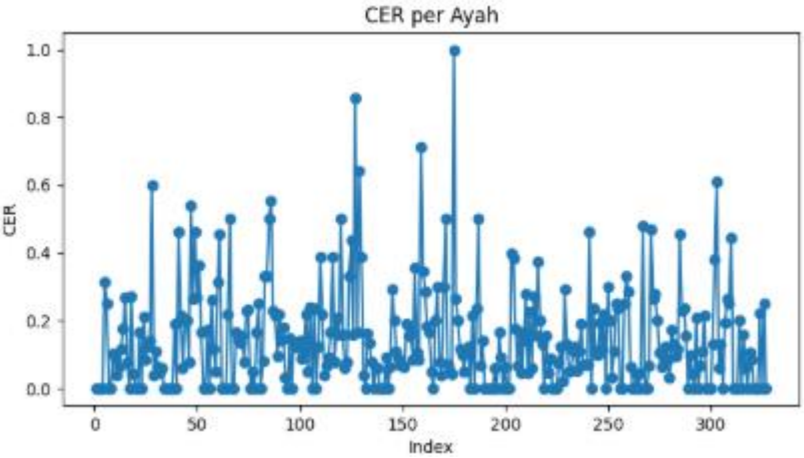

**Figure 6.** CER Per Father

**Table 2.** Example of a Result Table

| Name | Correct Text | Predicted Text | CER | Time/Father |
|---|---|---|---|---|
| 085001.mp3 | And the sky with the zodiac | And the sky with the zodiac | 0 | 7.797 |
| 085002.mp3 | And the promised day | And the promised day | 0 | 8.964 |
| 085003.mp3 | And to witness and praise | And to witness and praise | 0 | 9.139 |
| 085004.mp3 | Kill the owner of the flow | Kill the owner of the flow | 0 | 7.926 |
| 085005.mp3 | Fueled Fire | 7. Narzat Fuel | 2,2 | 9.457 |
| 085006.mp3 | As they sit on it | 10. When they sit on it | 00.25 | 8.527 |
| 085012.mp3 | Your Lord's Persecution Is Severe | If your Lord is so oppressed | 0,434 | 9.398 |
| 085013.mp3 | That's what starts and repeats. | That he began and promised | 0,1176 | 9.095 |
| 085014.mp3 | He was a forgiving and friendly person | It is a friendly fountain | 0,1764 | 7.637 |
| 085015.mp3 | The Noble Throne | Noble insult | 0,266 | 9.113 |

The study evaluated the Whisper model on 327 fathers of the last 30 surah of Juz 30, focusing on accuracy, processing time, and computational efficiency in a low-resource context. The results showed that the model took an average of 9.02 seconds to process one father, with a total processing duration of about 2950 seconds, or about 50 minutes for all fathers combined. The show was contextualized using data from https://www.knowledgequran.com/, which showed that a fluent reader typically completes one verse of the Qur'an in about 1 hour, an average of 3 minutes per page. For example, Surah An-Naba, which contains 40 verses and covers about 1.5 pages, will be recited by an eloquent individual in about 270 seconds. In contrast, the Whisper model copies the same surah in about 360 seconds based on its average speed — about 1.33 times slower than a live human read. This highlights a significant performance gap between the model's inference speed and real-time expectations, demonstrating Whisper's current limitations for interactive or real-time Qur'an applications.

Despite this latency, the model shows very efficient memory usage. On average, each father only needs 0.0023 MB of RAM, with minimal variance across all test samples. This ultra-low memory footprint makes the Whisper model particularly suitable for use in compute-constrained environments such as embedded systems or low-power mobile devices. Consistent use of RAM also supports its application in large-scale batch transcription setups where hardware efficiency is critical.

In terms of transcription accuracy, the model achieved an average Character Error Rate (CER) of 0.144 across all fathers. Most fathers were transcribed with perfect accuracy (CER = 0.0), while minorities experienced a higher error rate that contributed to the average. These differences suggest that while the model generally performs well, it is susceptible to certain types of recognition errors—such as confusion over similar-sounding phonemes or character insertion—in certain Arabic contexts. The lack of diacritics and tajweed symbols in the dataset likely reduces linguistic complexity, but a more nuanced analysis of errors will be needed to fully diagnose the cause of these errors.

## 4. CONCLUSION

This research shows that the Whisper model holds great potential for Qur'anic speech recognition, especially in low-resource settings, with good accuracy and high memory efficiency. With an average CER of 0.144 and RAM usage as low as 0.0023 MB per verse, this model can transcribe Qur'anic verses accurately using minimal hardware resources. However, the average processing time of more than 9 seconds per verse remains a significant barrier, as it is still too slow

for real-time applications, such as providing immediate feedback for Qur'an learners or automated correction tools. Therefore, future research should focus on optimizing processing time, through strategies such as model trimming, quantization, or using lighter model variants. Additionally, this study did not compare Whisper with other ASR models like VOSK or Wav2Vec2, which would offer a more comprehensive understanding of its performance. Experimenting with finer input segmentation—by breaking longer verses into smaller phonetic units—could speed up processing time and improve model responsiveness. Future work addressing these issues will be crucial for enabling real-time practical applications of Qur'anic speech recognition.

## REFERENCE

[1] A. Rifani, “BAHASA AL-QUR’AN SEBAGAI BAGIAN DALAM IJTIHADIYYAH,” 2019. [Online]. Available: https://jurnal.uin-antasari.ac.id/index.php/jils/issue/view/472

[2] N. Nurhanifah, “URGENSI PENDIDIKAN AL-QUR’AN: KAJIAN PROBLEMATIKA KETIDAKMAMPUAN MEMBACA AL-QUR’AN DAN SOLUSINYA,” *JUMPER: Journal of Educational Multidisciplinary Research*, vol. 2, no. 1, pp. 102–114, Jan. 2023, doi: 10.56921/jumper.v2i1.73.

[3] Zulfitria, “PERANAN PEMBELAJARAN TAHFIDZ AL-QURAN DALAMPENDIDIKAN KARAKTER DI SEKOLAH DASAR,” *Naturalistic: Jurnal Kajian Penelitian Pendidikan dan Pembelajaran 1*, no. 2, pp. 124–134, Apr. 2017.

[4] S. Susanto and M. A. Muhaidori, “The Role of Tahfidz Al-Quran Learning in Assisting Religious Studies,” *International Journal of Language and Ubiquitous Learning*, vol. 2, no. 2, Jul. 2024, doi: 10.70177/ijlul.v2i2.1150.

[5] N. M. Mustafa, Z. Mohd Zaki, K. A. Mohamad, M. Basri, and S. Ariffin, “Development and Alpha Testing of EzHifz Application: Al-Quran Memorization Tool,” *Advances in Human-Computer Interaction*, vol. 2021, 2021, doi: 10.1155/2021/5567001.

[6] R. A. Rajagede and R. P. Hastuti, “Al-Quran recitation verification for memorization test using Siamese LSTM network,” *Communications in Science and Technology*, vol. 6, no. 1, pp. 35–40, 2021, doi: 10.21924/CST.6.1.2021.344.

[7] D. Wang, X. Wang, and S. Lv, “An overview of end-to-end automatic speech recognition,” 2019, *MDPI AG*. doi: 10.3390/sym11081018.

[8] S. Alharbi *et al.*, “Automatic Speech Recognition: Systematic Literature Review,” 2021, *Institute of Electrical and Electronics Engineers Inc.* doi: 10.1109/ACCESS.2021.3112535.

[9] S. Al-Fadhli, H. Al-Harbi, and A. Cherif, “Speech Recognition Models for Holy Quran Recitation Based on Modern Approaches and Tajweed Rules: A Comprehensive Overview,” *IJACSA) International Journal of Advanced Computer Science and Applications*, vol. 14, no. 12, p. 2023, 2023, [Online]. Available: www.ijacsa.thesai.org

[10] M. Hadwan, H. A. Alsayadi, and S. AL-Hagree, “An End-to-End Transformer-Based Automatic Speech Recognition for Qur’an Reciters,” *Computers, Materials and Continua*, vol. 74, no. 2, pp. 3471–3487, 2023, doi: 10.32604/cmc.2023.033457.

[11] Y. He *et al.*, “Streaming End-to-end Speech Recognition For Mobile Devices,” Nov. 2018, [Online]. Available: http://arxiv.org/abs/1811.06621

[12] R. Prabhavalkar, T. Hori, T. N. Sainath, R. Schluter, and S. Watanabe, “End-to-End Speech Recognition: A Survey,” *IEEE/ACM Trans Audio Speech Lang Process*, vol. 32, pp. 325–351, 2024, doi: 10.1109/TASLP.2023.3328283.

[13] J. Li, “Recent Advances in End-to-End Automatic Speech Recognition,” Redmond, Feb. 2022. doi: 10.1561/116.00000050_supp.

[14] N. Sethiya and C. K. Maurya, “End-to-End Speech-to-Text Translation: A Survey,” Indore: Indian Institute of Technology, Jun. 2024.

[15] D. Ferdiansyah, C. Sri Kusuma Aditya, J. Raya Tlogomas No, K. Lowokwaru, K. Malang, and J. Timur, “Implementasi Automatic Speech Recognition Bacaan Al-Qur’an Menggunakan Metode Wav2Vec 2.0 dan OpenAI-Whisper,” *JURNAL TEKNIK ELEKTRO DAN KOMPUTER TRIAC*, vol. 11, no. 1, pp. 2615–7764, 2024, [Online]. Available: https://journal.trunojoyo.ac.id/triac

[16] A. Moustafa and S. A. Aly, “Towards an Efficient Voice Identification Using Wav2Vec2.0 and HuBERT Based on the Quran Reciters Dataset,” Nov. 2021, [Online]. Available: http://arxiv.org/abs/2111.06331

[17] A. Rahman, M. M. Kabir, M. F. Mridha, M. Alatiyyah, H. F. Alhasson, and S. S. Alharbi, “Arabic Speech Recognition: Advancement and Challenges,” *IEEE Access*, vol. 12, pp. 39689–39716, 2024, doi: 10.1109/ACCESS.2024.3376237.

[18] A. A. Abdelhamid, H. A. Alsayadi, and I. Hegazy, “End-to-End Arabic Speech Recognition: A Review,” Oct. 2020. [Online]. Available: https://www.researchgate.net/publication/344799361

[19] A. Purbasari, F. R. Rinawan, A. Zulianto, A. I. Susanti, and H. Komara, “CRISP-DM for Data Quality Improvement to Support Machine Learning of Stunting Prediction in Infants and Toddlers,” in *Proceedings - 2021 8th International Conference on Advanced Informatics: Concepts, Theory, and Application, ICAICTA 2021*, Institute of Electrical and Electronics Engineers Inc., 2021. doi: 10.1109/ICAICTA53211.2021.9640294.

[20] J. Brzozowska, J. Pizoń, G. Baytikenova, A. Gola, A. Zakimova, and K. Piotrowska, “DATA ENGINEERING IN CRISP-DM PROCESS PRODUCTION DATA – CASE STUDY,” *Applied Computer Science*, vol. 19, no. 3, pp. 83–95, 2023, doi: 10.35784/acs-2023-26.

[21] C. Schröer, F. Kruse, and J. M. Gómez, “A systematic literature review on applying CRISP-DM process model,” in *Procedia Computer Science*, Elsevier B.V., 2021, pp. 526–534. doi: 10.1016/j.procs.2021.01.199.

[22] A. Rianti *et al.*, “CRISP-DM: Metodologi Proyek Data Science,” Prosiding Seminar Nasional Teknologi Informasi dan Bisnis (SENATIB), 2023.

[23] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, “Robust Speech Recognition via Large-Scale Weak Supervision,” Dec. 2022, [Online]. Available: http://arxiv.org/abs/2212.04356

[24] S. Alrumiah and A. Alshargabi, “A Deep Diacritics-Based Recognition Model for Arab,” *IEEE Access*, vol. 10, 2022.

[25] S. Fradj, “Speaker Recognition and Automatic Speech Recognition ,A personal project exploring methods and techniques in Speaker Recognition and Automatic Speech Recognition,” *Tunis Business School*, Mar. 2025, doi: 10.5281/zenodo.15102949.

[26] Q. A. Obaidah *et al.*, "A New Benchmark for Evaluating Automatic Speech Recognition in the Arabic Call Domain," Mar. 2024, [Online]. Available: http://arxiv.org/abs/2403.04280

[27] Thennal, D. Gopinath, J. James, and M. Ashraf, "Advocating Character Error Rate for Multilingual ASR Evaluation," Cornell University, Oct. 2024. doi: https://doi.org/10.48550/arXiv.2410.07400.

[28] A. A. Sodhar, T. H. Ansari, and A. Q. Channa, "Introduction and history of Qur'an recitation," *Al Khadim Research Journal of Islamic Culture and Civilization*, vol. V, no. 3, pp. 183–205, 2024, [Online]. Available: https://www.arjicc.com

[29] A. N. Farahdiba *et al.*, "Bringing the Qur'an to life: Teaching students the art of reciting the Qur'an," *Jurnal Pembelajaran Pemberdayaan Masyarakat (JP2M)*, vol. 5, no. 2, pp. 295–305, Jun. 2024, doi: 10.33474/jp2m.v5i2.21704.

[30] A. Andreyev, "Quantization for OpenAI's Whisper Models: A Comparative Analysis," 2025.

[31] C. Graham and N. Roll, "Evaluating OpenAI's Whisper ASR: Performance analysis across diverse accents and speaker traits," *JASA Express Lett*, vol. 4, no. 2, Feb. 2024, doi: 10.1121/10.0024876.

[32] Y. Liu, X. Yang, and D. Qu, "Exploration of Whisper fine-tuning strategies for low-resource ASR," *EURASIP J Audio Speech Music Process*, vol. 2024, no. 1, Dec. 2024, doi: 10.1186/s13636-024-00349-3.

[33] N. San *et al.*, "Predicting positive transfer for improved low-resource speech recognition using acoustic pseudo-tokens," Feb. 2024, [Online]. Available: http://arxiv.org/abs/2402.02302

[34] A. Waheed, H. Atwany, R. Singh, and B. Raj, "On the Robust Approximation of ASR Metrics," 2025.