Speech Recognition Models for Holy Quran Recitation Based on Modern Approaches and Tajweed Rules: A Comprehensive Overview

Sumayya Al-Fadhli[1], Hajar Al-Harbi[2], Asma Cherif[3]
Department of Computer Science, King Abdulaziz University, Jeddah, Saudi Arabia[1,2]
Department of Computer Science-Adham University College, Umm Al-Qura University, Makkah, Saudi Arabia[1]
Department of Information Technology, King Abdulaziz University, Jeddah, Saudi Arabia[3]
Center of Excellent in Smart Environment Research, King Abdulaziz University, Jeddah, Saudi Arabia[3]

***Abstract*—Speech is considered the most natural way to communicate with people. The purpose of speech recognition technology is to allow machines to recognize and understand human speech, enabling them to take action based on the spoken words. Speech recognition is especially useful in educational fields, as it can provide powerful automatic correction for language learning purposes. In the context of learning the Quran, it is essential for every Muslim to recite it correctly. Traditionally, this involves an expert *gari* who listens to the student's recitation, identifies any mistakes, and provides appropriate corrections. While effective, this method is time-consuming. To address this challenge, apps that help students fix their recitation of the Holy Quran are becoming increasingly popular. However, these apps require a robust and error-free speech recognition model. While recent advancements in speech recognition have produced highly accurate results for written and spoken Arabic and non-Arabic speech recognition, the field of Holy Quran speech recognition is still in its early stages. Therefore, this paper aims to provide a comprehensive literature review of the existing research in the field of Holy Quran speech recognition. Its goal is to identify the limitations of current works, determine future research directions, and highlight important research in the fields of spoken and written languages.**

Keywords*—Speech recognition; acoustic models; language model; neural network; deep learning; quran recitation

I. Introduction

Speech is the most natural way to communicate with people [7]. Designing a machine that mimics human behavior, including speaking naturally and responding correctly to spoken language, has puzzled engineers and scientists for centuries [51]. Automatic speech recognition (ASR) refers to the computational process of transforming acoustic speech signals into written words or other linguistic units through dedicated algorithms [28], [7]. The goal of ASR is to enable machines to interpret and respond to spoken language [4]. ASR involves the capability of a machine to accurately recognize speech, convert it into text, and take appropriate actions based on human instructions [7]. In particular, speech recognition is useful in educational fields as it allows for the building of powerful automatic correctors for language learning purposes. As [41], they build a model of English pronunciation learning for Chinese learners.

The researchers have made significant contributions to speech processing in various languages spoken worldwide. There are three classes in Arabic, which has approximately 420 million speakers [47]. The primary class taught in schools is Modern Standard Arabic (MSA), which adheres to the grammatical rules of the Arabic language. The second class is Arabic Dialect (AD), which represents the everyday spoken language of native Arabic speakers, varying across countries and regions. The third class is classical Arabic (CA), the language used in the Holy Quran, which has been renowned globally for centuries. CA is known for its extensive grammar and vocabulary, as well as its unique recitation guidelines [15], [24].

Recently, the use of speech recognition in the Quranic recitation field has emerged as an important research direction. Indeed, there are more than two billion Muslims in the world [2]. Muslims generally strive to learn the precise recitation of CA and adhere to certain rules known as *Tajweed* in order to recite the Holy Quran accurately. Learning these rules is very important for all Muslims to master the recitation of the Holy Quran [15]. Consequently, building accurate Holy Quran Speech Recognition (HQSR) models represents a significant research outcome for all Muslims.

Teaching the correct recitation of the Quran is essential for every Muslim. Learning Quran recitation usually depends on an expert, also known as *gari*, who listens to the student's recitation, determines recitation mistakes, and instructs the student with the appropriate correction. This way of learning is very effective, but it's time-consuming because the teacher needs to correct the errors of every student independently. For this reason, the apps that help students fix their recitation of the Holy Quran are beneficial and essential, but these apps need a robust and error-free speech recognition model. Despite conducting several research studies in this area, researchers have not yet achieved the optimal solution for recognizing speech in the Holy Quran. Though recent models have been applied to written and spoken Arabic and non-Arabic speech recognition and produced highly accurate results, Quran speech recognition is still in its early stages. Therefore, this paper aims to propose a comprehensive literature review of the works in the field of Holy Quran speech recognition and shed light on some important research in the field of spoken and written languages.

The main motivation for our research is as follows:

1) Though the Holy Quran represents an essential book for all Muslims, current models for Holy Quran speech recognition have low accuracy or do not cover all chapters (i.e., rely on small datasets).
2) Some people find it difficult to attend Quran learning courses or retrieve their memorization in front of the teacher. Many individuals struggle to retain the Quran due to fear. Thus, building a professional app for Quran learning is important to help them retain the Quran in their home.
3) Quran memorization requires a continuous review process, which is time-consuming. Thus, it is hard for Quran teachers to listen and validate long recitations for many students.
4) Some people prefer reading what they memorize, especially in night prayer (i.e., without reading from the *Mushaf* to not lose their submission in prayer). However, they can easily make mistakes. An automatic corrector can assist Muslims in their prayers.
5) Some non-Arabic countries, mainly those with a minority of Muslims, do not have enough qualified teachers to teach the Holy Quran.

However, research in the field of HQSR is still in its early stages. Indeed, recognizing individual words is easy, but the challenge is recognizing continuous recitation [7] and detecting erroneous recitation and violations of tajweed rules. In the realm of speech recognition systems (see Fig. 1), various factors, such as speaker dependency, vocabulary size, and noisy environments, can significantly impact their performance. Recognition performance increases with limited vocabulary and reciter-dependent conditions while using broad vocabulary and reciter-independent scenarios; performance can decrease significantly [7]. Besides, most research developments focus on one or a few chapters or a few tajweed rules. Also, existing works in HQSR suffer from the lack of large datasets used. Finally, current works use traditional techniques and do not investigate end-to-end learning.

The critical objective of this research is to use machine learning for Holy Quran recitation. Our main contributions are to provide a thorough literature review to find the most important issues that need more investigation in the field of Holy Quran speech recognition. Moreover, our study summarizes some important and informative papers in the Arabic and non-Arabic languages fields and recent papers in the HQSR field and provides a taxonomy for speech recognition and HQSR.

Various machine learning algorithms could be used in speech recognition, including Dynamic Time Warping (DTW), Hidden Markov Models (HMM), and Artificial Neural Networks (ANN) [51].

In the context of Arabic ASR, many algorithms were used, such as recurrent neural networks (RNN), long short-term memory (LSTM), which is a particular case of RNN, and connectionist temporal classification (CTC) [4].

The remaining parts of this paper are structured in the following way: Section II discusses some important research in written and spoken languages speech recognition. Section III discusses the recent papers in the field of Holy Quran Speech Recognition. Next, Section IV discusses some of the research directions in the field of Holy Quran Speech Recognition. Finally, Section V concludes the paper.

II. Speech Recognition for Written & Spoken Languages

In this section, we highlight some speech recognition solutions that produced impressive results in Arabic and non-Arabic languages. These solutions may be categorized as traditional speech recognition (either with deep learning architecture or without deep learning) or as an end-to-end-based speech recognition solution.

A. Traditional Models

Fig. 2 shows that traditional ASR systems are made up of three separate parts: the acoustic model, the pronunciation model, and the language model [54]. The Acoustic Model (AM) assesses the likelihood of acoustic units such as phonemes, graphemes, or sub-word units [13]. In contrast, the Language Model (LM) evaluates the likelihood of word sequences. By integrating linguistic knowledge derived from extensive text collections, language models improve the precision of acoustic models. These models use the acquired syntactic and semantic rules to re-evaluate the hypotheses generated by the acoustic model. The process of mapping a series of phonemes to words is done by the Pronunciation Dictionary (PD), and it aligns the phonetic transcriptions produced by the AM system with the unprocessed text used in language models. The training of these three components is done individually, and then they are merged together to form a search graph by utilizing finite-state transducers (FSTs). Feature Extraction (FE) takes input speech as input, produces the essential features, and then sends these features to the decoder. Following that, the decoder produces lattices, which are then evaluated and ordered to generate the desired sequences of words.

The acoustic model can be modeled using HMMs [32] and Gaussian Mixture Models (GMMs) [61]. It is worth noting that recent ASR models have replaced the use of GMMs in the acoustic model with deep neural networks (DNNs) [30]. These are referred to as hybrid HMM-DNN and are widely used as competitive ASR models. Also, some research replaced GMMs with Bidirectional Long-Short Term Memory (BLSTM) [50], while some other studies replaced HMM with another classification method such as Support Vector Machine (SVM) [12], [36], Linear Discriminant Analysis (LDA) combined with Quadratic Discriminant Analysis (QDA) [33], Convolutional Neural Networks (CNNs) and SVM [40], and Hidden Semi-Markov Model (HSMM) [34], etc.

Some research has been suggested to address speech recognition for Arabic and non-Arabic languages.

Arabic Language. In their study, [39] introduced a novel approach that combines three distinct training systems for speech recognition. Four-gram language model re-scoring, system combination with minimum Bayes risk decoding, and lattice-free maximum mutual information are a few of these groups. They achieved significant progress, with a word error rate of 42.25% on the Multi-Genre Broadcast (MGB-3) Arabic

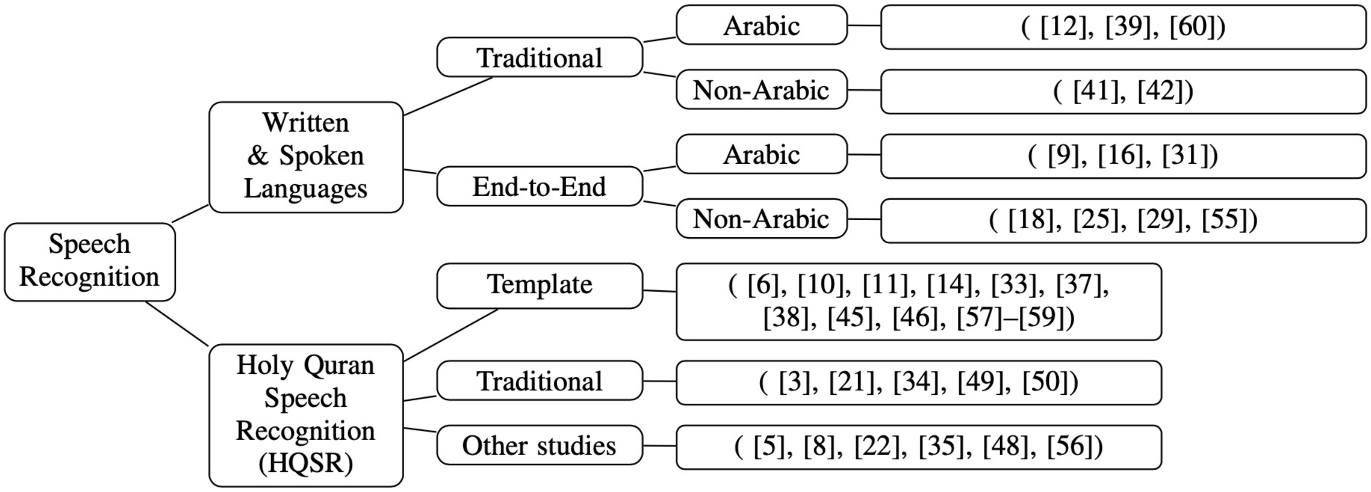

Fig. 1. Speech recognition taxonomy.

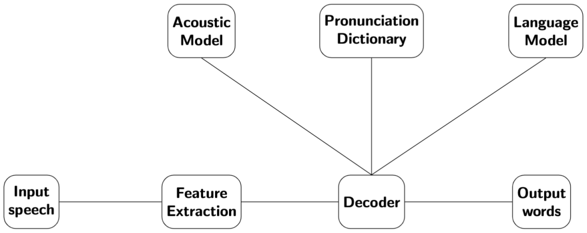

Fig. 2. Traditional ASR Pipeline ([54]).

development set. They got this result by using a 4-gram rescoring strategy for a chain BLSTM system. This system did better than a DNN system that had a word error rate of 65.44%.

In [60], the authors presented a comprehensive framework for Arabic speech recognition. To turn sequences of Mel Frequency Cepstral Coefficients (MFCC) and Filter Bank (FB) features into fixed-size vectors, they used recurrent LSTM or GRU architectures. They then fed these vectors into a multilayer perceptron network (MLP) to perform classification and recognition tasks. The researchers evaluated their system using two different databases: one for spoken-digit recognition and another for spoken TV commands. However, a limitation of their work is the absence of datasets that incorporate recorded speech signals in noisy, realistic environments.

In [12], they presented a speech recognition system for the Arabic language. The system aimed to evaluate three feature extraction algorithms: MFCC, Power Normalized Cepstral Coefficients (PNCC), and Modified Group Delay Function (ModGDF). We performed the classification process using an SVM. The results indicated that PNCC was the most effective algorithm, while ModGDF achieved moderate accuracy. PNCC and ModGDF outperformed MFCC in terms of precision. PNCC achieved an accuracy rate of 93% to 97%, ModGDF achieved 90%, and MFCC achieved 88%.

Non-Arabic languages. The authors of [42] presented the design of Kaldi, a speech recognition toolkit that is freely available and open-source. The highly permissive Apache License v2.0, under which Kaldi is released, enables extensive usage. Kaldi provides a robust speech recognition system that utilizes finite-state transducers and is built on the OpenFst library. The toolkit provides comprehensive documentation and scripts that make it easier to build comprehensive recognition systems. Kaldi is coded in C++, and its core library offers a range of functionalities, including phonetic-context modeling, acoustic modeling using subspace Gaussian mixture models (SGMM), standard Gaussian mixture models, and linear and affine transforms.

A speech-learning system for the English language was developed and implemented by [41]. It utilized a speech recognition technique based on HMM to decode speech using the Viterbi algorithm and determine the recognition score through posterior probability. The system achieved an average recognition rate of 94%. Its purpose was to help English learners assess pronunciation accuracy during verbal practice and identify different types of errors. By engaging in systematic practice with this system, users can significantly enhance their listening and speaking skills. The system provides real-time feedback on oral pronunciation accuracy, error correction reports, and allows for repeated practice to facilitate effective training.

Table I summarizes traditional speech recognition techniques for Arabic and non-Arabic languages.

B. End-to-End-based Speech Recognition Models

The purpose of the end-to-end (E2E) system is to directly transform a series of acoustic features into a corresponding series of graphemes or words. This approach greatly simplifies traditional speech recognition methods by eliminating the need for manual labeling of information in the neural network. Instead, the E2E system automatically learns language and pronunciation information, as depicted in Fig. 3.

End-to-end speech recognition systems typically rely on an encoder-decoder framework. According to studies [17], [13], this architecture takes an audio file as input and processes it through a series of convolution layers to generate a condensed

TABLE I. Summary of Traditional Speech Recognition Techniques in Written and Spoken Languages

Ref	Lang.	Main idea	FE	Classification	LM	AM
[12]	Arabic	Study three feature extraction methods, MFCC, PNCC, and ModGDF for the development of an ASR System in Arabic.	MFCC, PNCC, and ModGDF	SVM	-	-
[60]	Arabic	Prsent an approach based on RNN to process sequences of variable lengths of MFCCs, FBs and delta-delta features of the different spoken digits/commands.	MFCC (static and dynamic features), and the FB coefficients.	LSTM and Neural Network (MultiLayer Perceptron: MLP) classifier	-	BiLSTM model
[39]	Arabic	Improve Hybrid ASR for MGB & Al-Jazeera speech data.	MFCCs features	Hybrid ASR using TDNN-LSTM & Bi-directional prioritized grid LSTM (BPGLSTM)	n-gram LM	Hybrid ASR using TDNN-LSTM & BPGLSTM
[42]	Non-Arabic	Build a new open source toolkit for Conventional speech recognition from scratch called Kaldi toolkit.	MFCC features	GMM-DNN-HMM	bigram	DNN-HMM
[41]	Non-Arabic	Build a leight-weight speech recognition using GMM-HMM, to learn English language using HMM based speech recognition for Chinese speakers.	MFCC features	HMM based	n-gram	GMM-HMM

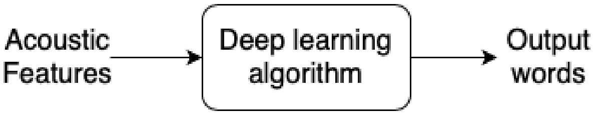

Fig. 3. End-to-End ASR Pipeline ([54]).

vector. The decoder then uses this vector to generate a character sequence. Researchers can use different objective functions, such as CTC [23], ASG [20], LF-MMI [25], sequence-to-sequence [18], transduction [44], and differentiable decoding [19], to optimize the end-to-end ASR [13]. Researchers have also explored different neural network architectures, including ResNet [27], TDS [26], and Transformer [52]. Additionally, integrating an external language model has been shown to improve the overall performance of the system.

Recently, some research has been suggested for both Arabic and non-Arabic speech recognition using end-to-end models. In what follows, we discuss and summarize these solutions.

Arabic Language. In [9], the researchers introduced the first comprehensive approach to building an Arabic speech-to-text transcription system. They utilized lexicon-free RNNs and the CTC objective function to achieve this. The system consisted of three main components: a BDRNN acoustic model, a language model, and a character-based decoder. Unlike word-level decoders, their decoder did not rely on a lexicon during the transcription process. The RNN acoustic and language model successfully distinguished between characters with the same accent but different writing styles. The researchers evaluated the model using a 1200-hour corpus of Aljazeera multi-genre broadcast programs, resulting in a 12.03% word error rate for non-overlapped speech. It is important to note that deep learning techniques were only used in the feature extraction phase.

In [16], the authors proposed a robust diacritized ASR system using both traditional ASR and end-to-end ASR techniques. They trained and tested their models on the Standard Arabic Single Speaker Corpus (SASSC) with diacritized text data, using MFCCs and FB for feature extraction. The ASR speech recognition system incorporated a total of eight models, comprising four GMM models, two SGMM models, and two DNN models. We constructed these models using the KALDI toolkit and performed language modeling using CMU-CLMTK. The best achieved word error rate (WER) among these models was 33.72% using DNN-MPE. Additionally, the authors proposed an end-to-end approach for diacritized Arabic ASR, employing joint CTC-attention and CNN-LSTM attention methods. The CNN-LSTM with attention method outperformed the others, achieving a character error rate (CER) of 5.66% and a WER of 28.48%. This method resulted in a significant reduction in WER compared to both the traditional ASR and joint CTC-attention method, by 5.24% and 2.62%, respectively.

Researchers conducted a comprehensive comparison on Arabic language and its dialects using different ASR approaches in [31]. The researchers collected a new evaluation set comprising news reports, conversational speech, and various datasets to ensure unbiased analysis. They extensively analyzed the errors and compared the ASR system's performance with that of expert linguists and native speakers. While the machine ASR system showed better performance than the native speaker, there was still an average WER gap of 3.5% compared to expert linguists in raw Arabic transcription. The proposed end-to-end transformer model outperformed prior state-of-the-art systems on MGB2, MGB3, and MGB5 datasets, achieving new state-of-the-art performances of 12.5%, 27.5%, and 33.8%, respectively.

Non-Arabic Languages. Researchers introduced ESPnet, a novel open-source platform for end-to-end speech processing, in [55]. ESPnet leverages dynamic neural network toolkits like Chainer and PyTorch, serving as the primary deep learning engine. This platform simplifies the training and recognition processes of the entire ASR pipeline. The ESPnet uses the same feature extraction/format, data processing, and scheme style as the Kaldi ASR toolkit. This gives researchers a complete way to test speech recognition and other speech processing techniques. The test results show that ESPnet does a good job with ASR and is about as efficient as the most advanced HMM/DNN systems that use traditional setups. Notably, ESPnet has made significant advancements, including the incorporation of multi-GPU functionality (up to 5 GPUs). In just 26 hours, ESPnet successfully completed the training of 581 hours of the CSJ task.

In [25], the authors described a simple HMM-based end-to-end method for ASR and tested how well it worked on well-known large-vocabulary speech recognition tasks, specifically the Switchboard and Wall Street Journal (WSJ) corpora. The authors trained the acoustic model used in this approach without the need for initial alignments, prior training, pre-estimation, or transition training, making it entirely neural except for the decoding/LM part. The proposed method surpassed other end-to-end methods in similar setups, particularly when dealing with small databases. By employing a comprehensive biphone modeling approach, the researchers achieved results almost comparable to regular LF-MMI training.

In [18], the researchers introduced a new attention-based model called Listen, Attend, and Spell (LAS) for sequence-to-sequence speech recognition. LAS combines the sound, pronunciation, and language model parts of regular ASR systems into a single neural network, so there's no need for a separate dictionary or text normalization. The researchers compared LAS with a hybrid HMM-LSTM system and found that LAS achieved a WER of 5.6%, outperforming the hybrid system's WER of 6.7%. In a dictation task, LAS achieved a WER of 4.1%, while the hybrid system achieved a WER of 5%.

The study [29] introduced a new strategy to address multilingual ASR speech recognition, specifically in the context of code-switching speech. The researchers employed three techniques to achieve this. The researchers decoded the speech by utilizing a global language model constructed from multilingual text. Their system used a multigraph approach along with weighted finite-state transducers (WFST), which let them switch between languages while decoding by using a closure operation. The output of this process was a bilingual or multilingual text based on the input audio. Secondly, they employed a robust transformer system for speech decoding. They found that WFST decoding was particularly suitable for inter-sentential code-switching datasets among the techniques used.

Table II summarizes end-to-end speech recognition techniques for Arabic and non-Arabic languages.

In the following, we compare the end-to-end architecture mentioned in the previous works with the baseline traditional techniques on the same datasets as indicated in the previous reviewed studies. As we see in Table III, the end-to-end architecture outperforms the hybrid architecture in all the studies mentioned in the table, except [25].

III. Holy Quran Speech Recognition (HQSR)

This section summarizes the recent studies that concern the HQSR. It also presents the used techniques and the performance of the proposed solutions. Moreover, it determines the gap and limitations in the current research on the HQSR. We classified the current studies of HQSR into three categories: template-based speech recognition, traditional-based speech recognition, and other HQSR studies.

A. Template Based Speech Recognition

This section summarizes the papers that follow template-based speech recognition, which is an old style of speech recognition that relied only on FE, classification, and matching techniques (i.e., it didn't have acoustic, lexical, or language models).

In [46], the authors provide a deep learning model utilizing a dataset of seven famous reciters and CNNs. They employ MFCCs to extract and assess data from audio sources. Their provided model achieved 99.66% accuracy.

In [6], the authors highlight the key distinctions between a basic ASR system and an ASR-based language tutor specifically designed for Quran memorization. They demonstrate that ASR techniques alone are not sufficient for an intelligent Quran tutor and propose modifications to enhance its capabilities. To support their claims, the researchers utilize data from Sūrat Al-Nass. However, one of the major obstacles to developing a Quran tutor is the absence of a comprehensive dataset containing both correct and incorrect recitations, which is necessary for conducting meaningful experiments in this domain.

In [37], researchers proposed an online verification system for Quran verses to ensure the integrity and authenticity of the Quran. They gathered data from ten expert Qari who recited Surat Al-Nass ten times correctly and ten times with various types of mistakes (e.g., Tajweed, Makhraj, missing words). Unlike modern techniques, this study did not utilize acoustic, lexical, or language models. Instead, it relied on MFCC for feature extraction and HMMs for recognition and matching. However, the study did not provide any testing results.

In [14], the authors center on the examination and identification of classical Arabic vocal phonemes, specifically vowels, through the utilization of HMM. They aim to tackle the issue of semantic changes that can occur due to variations in vowel durations (short or long) in Arabic. To investigate this, they examine three chapters (Alfateha, Albaqarah, and Alshuraa) from the Holy Quran. Their findings demonstrate an impressive overall accuracy rate of 87.60% without the utilization of a specified language model.

In [58], researchers developed a speech recognition system that utilizes MFCC for feature extraction and HMM for classification. The system focuses on recognizing and identifying the rules of Iqlab on Qira'at of Warsh. It uses a database of expert teachers' rules to compare and report any mismatches in the iqlab rules for specific verses. The system achieved a 70% accuracy in correctly spelling words with the correct rules from the database, a 50% accuracy for words with incorrect rules from the database, and a 40% accuracy for new words not included in the training database.

In [57], the researchers presented an interactive Tajweed system that assists in verifying the appropriate Imaalah Checking rule for Warsh recitation. This system utilizes an auto-

TABLE II. SUMMARY OF END-TO-END BASED SPEECH RECOGNITION IN WRITTEN AND SPOKEN LANGUAGES

Ref	**Lang.**	**Main idea**	**FE**	**E2E DL**	**LM**	**AM**
[29]	Arabic &Non-Arabic	A new strategy for multilingual ASR speech recognition. Implementation of three strategies to identify code-switching speech.	MFCCs and FB	Transformer based E2E Architecture	n-gram	TDNN & Transformer based E2E architecture
[31]	Arabic	A thorough examination to compare the E2E transformer ASR, the modular HMM-DNN ASR, and HSR.	MFCCs and Mel-spectrogram.	E2E transformer with hybrid (CTC+Attention)	LSTM and transformer-based language model (TLM)	combining a TDNN with LSTM layers.
[16]	Arabic	Build a robust diacritised Arabic ASR	MFCCs and the log Mel-Scale Filter Bank energies.	using joint CTC attention and using CNN-LSTM with attention	built by CMUCLMTK tool based on the 3-g and trained on RNN-LM.	KALDI, ESPnet, and Espresso
[9]	Arabic	E2E model for Arabic speech-to text transcription system using the lexicon free RNNs and CTC objective function based on Stanford CTC source code.	FB	BDRNNs	n-gram	TDNN-LSTM & BDRNNs
[25]	Non-Arabic	E2E training of AM using the LF-MMI objective function in the context of HMMs	MFCC	TDNN-LSTM	n-gram & RNN	E2E LF-MMI
[18]	Non-Arabic	Improving the performance of LAS which is a novel technique in ASR research.	FB	LAS ASR	5-gram	LAS ASR
[55]	Non-Arabic	Proposed purely E2E speech recognition open source framework called ESPnet toolkit.	MFCC & FB	E2E SR framework	RNN LM	CTC-objective fucnction

TABLE III. END-TO-END BASED SPEECH RECOGNITION PERFORMANCE COMPARED TO BASELINE TRADITIONAL TECHNIQUES

Ref.	**Lang.**	**Dataset**	**Baseline Traditional Tech.**		**End-to-End Tech.**	
			Model	**WER/CER**	**Architecture**	**WER/CER**
[29]	Arabic& non-Arabic	Arabic MGB2 &English TEDLIUM-3 & ES-CWA Corpus	Hybrid ASR	9.8%	E2E-Transformer	**8.29**%
[31]	Arabic	MGB2, and (Hidden Test (HT))	HMM-DNN	HT:15.9% MGB2: 15.8%	E2E-T (CTC + Attention)	HT:**12.6**% MGB2: **12.5**%
[16]	Arabic	Standard Arabic Single Speaker Corpus (SASSC)	Kaldi toolkit using DNN, MPE, and SGMM	33.72%	CNN-LSTM with attention using Espresso toolkit	**28.48**%
[9]	Arabic	8 hours Aljazeera corpus 1200 hours of TV Aljazeera corpus	TDNN-LSTM-BLSTM	14.7%	BDRNN with CTC objective function	**12.03**%
[55]	non-Arabic	Corpus of Spontaneous Japanese (CSJ)	HMM/DNN (Kaldi nnet1)	eval1:9.0% eval2:7.2% eval3:9.6%	ESPnet (i.e., VGG2-BLSTM, char-RNNLM, and joint decoding)	eval1:**8.7**% eval2:**6.2**% eval3:**6.9**%
[25]	non-Arabic	Switchboard And WSJ	Regular LF-MMI	Switchboard: **9.1**% WSJ: **2.8**%	E2E-LF-MMI	Switchboard: 9.6% WSJ:3.0%
[18]	non-Arabic	12,500 hour training set consisting of 15 million English utterances	hybrid HMM-LSTM	6.7% dictation task 5%	LAS end-to-end model	**5.6**% dictation task **4.1**%

matic speech recognition system with MFCC as the feature extraction technique and HMM as the classification method. The researchers conducted experiments using fifteen speech samples. The results showed that the system achieved a 60% accuracy rate for identifying the Imaalah rule based on the Warsh narration in the training data.

Researchers developed a system in [10] that identifies the Ahkam Al-Tajweed in a specific audio recording of Quranic recitation. The study focused on eight rules: "EdgamMeem" (one rule), "EkhfaaMeem" (one rule), "Ahkam Lam" in 'Allah' Term (two rules), and "Edgam Noon" (four rules). The classification problem involved 16 classes, covering the entire Holy Quran for verses that contained the eight rules. The system utilized various feature extraction techniques, including traditional methods like MFCC and LPC as well as newer methods like CDBN. Classifiers such as SVM and RF were employed, with the best accuracy of 96.4% achieved using SVM for classification and features extracted through MFCC, WPD, HMM-SPL, and CDBN.

In [59], authors developed a speech recognition system that can accurately differentiate between different types of Madd (elongated tone) and Qira'at (method of recitation) related to Madd. The system utilized MFCC as a feature extraction technique and HMM as a classification method. The focus of the study was on two specific types of Madd: greater connective prolongation and exchange prolongation rules for Hafss and Warsh. We collected a total of sixty data samples for analysis. The results showed that the accuracy of identifying the exchange prolongation rule was 60% for Warsh and 50% for Hafss. Additionally, the accuracy for identifying the greater connective prolongation rule was 40% for Warsh and 70% for Hafss.

Researchers developed an automated self-learning system in [33] to support the traditional method of teaching and learning Quran. The system aimed to classify the characteristics of Quranic letters. The study collected audio data from 30 participants, including 19 males and 11 females. The participants recited each sukoon alphabet once without repetition. The system used the Sukoon alphabet from the Quran to provide a description of the Makhraj (point of articulation) and Sifaat (characteristics) of each letter. The study successfully identified and classified the characteristic features of the alphabet, specifically in terms of learning (Al-Inhiraf) and repetition (Al-Takrir). The results showed that using QDA with all 19 features achieved the highest accuracy, with 82.1% for leaning (Al-Inhiraf) and 95.8% for repetition (Al-Takrir) characteristics.

The researchers conducted the study with the aim of creating a comprehensive system that accurately recognizes and determines the correct pronunciation of different Tajweed rules in audio. To achieve this, the researchers employed 70 filter banks as a feature extraction technique and utilized SVM as the classification method. The study focused on four specific rules, namely Ekhfaa Meem, Edgham Meem, Takhfeef Lam, and Tarqeeq Lam. The study utilized a dataset of 80 records, comprising a total of 657 recordings, encompassing both correct and incorrect recitations for each rule. They tested the models in the system against 30% of the recorded data and achieved a validation accuracy of 99%. [38] developed a recognition model to identify the "Qira'ah" from the corresponding Holy Quran acoustic wave. The study utilized MFCC as the feature extraction technique and SVM as the classification method. With a dataset of 258 wave files for 10 "Qira'ah" and including various reciters, the SVM accuracy achieved approximately 96%, while the accuracy of the ANN was 62%.

In another study, [45] proposed a system that automates the process of checking Tajweed for children who are learning the Quran. The system used the MFCC algorithm to extract the input speech signal and the HMM algorithm to compare children's recitation with the recitation stored in the database. However, this project focused solely on Surah Al-Fatihah and did not provide any testing results.

Table IV summarizes the previous studies of template-based speech recognition in the HQSR field.

B. Traditional Based Speech Recognition

This section summarizes the papers that follow traditional speech recognition as described in Section II-A.

The researchers aimed to develop a precise Arabic recognizer for educational purposes in the study conducted by [34]. They implemented an HSMM model with the primary objective of improving the durational behavior of the traditional HMM model. To achieve this, they utilized a corpus consisting of recordings from 10 reciters, totaling over 487 minutes of speech. They meticulously segmented the corpus at three levels: phoneme, allophone, and words, with precise time boundaries. They obtained the recordings by reciting the Holy Quran, covering all the essential Arabic sounds. As a result of their work, the recognition accuracy saw an improvement of approximately 1.5%.

The researchers constructed an acoustic model using the Carnegie Melon University (CMU) Sphinx trainer [21]. The CMU Sphinx trainer utilized recordings from 39 different reciters and 49 chapters (surah) to build a robust framework for continuous speech recognition. The acoustic model achieved an impressive WER of approximately 15%, showcasing its accuracy and effectiveness.

In their research, [50] utilized data from everyayah.com, a website that provides open-access Quran recitations by numerous professional reciters, including Sheikh [1]. They adopted a deep learning approach to train an acoustic model for Quranic speech recognition. The study focused on 13 different reciters and concluded that the hybrid HMM-BLSTM method outperformed the HMM-GMM method in terms of speech recognition accuracy. The baseline models (HMM-GMM) achieved an average WER of 18.39%. In contrast, the acoustic model using Hybrid HMM-BLSTM achieved significantly better results, with an average WER of 4.63% in the same testing scenario.

The researchers utilized the KALDI toolkit to create and assess a speaker-independent continuous speech recognizer specifically designed for Holy Quran recitations in a study [49]. The researchers successfully developed a large-vocabulary system capable of recognizing and analyzing Quranic recitations. They use 32 recitations for Chapter 20 (Sūrat Taha), according to Hafs from the A'asim narration. The most effective experimental configuration involves utilizing Time Delay Neural Networks (TDNN) with a sub-sampling

TABLE IV. SUMMARY OF HQSR TEMPLATE-BASED SPEECH RECOGNITION

Ref#	Main Idea	Dataset	FE	ML algo.	Pros	Cons
[45]	Automated tajweed Checking System for Children.	Surah Al-Fatihah. Ten respondents' recitation for testing purposes. One audio of correct recitation is used for comparison with the respondents' audio	MFCC	HMM	-	very small data set & no testing result
[46]	The objective of this research was to differentiate between reliable and unethical Qur'anic reciters.	Seven well-known Qur'anic reciters have been gathered into a dataset. On an audio file, each reciter recited the Quran's surahs for eighty minutes.	MFCC	CNN	The proposed system stages are well-organized and easily understandable.	-
[38]	A recognition model for the "Qira'ah" from the corresponding Holy Quran acoustic wave.	The corpus contains 258 wave files labeled based on the "Qira'ah" (they consider 10 "Qira'ah") .	MFCC	SVM	good accuracy (96.12 %)	-
[11]	A system for recognizing/correcting the different rules of Tajweed in an audio.	Almost 80 records for each rule name and type, a total of 657 recordings of 4 different rules.	FBs	SVM	Good process for data collection.	Consider only 4 rules
[33]	Features identification and classification of alphabet (ro) in Leaning (Al-Inhiraf) and Repetition (AlTakrir) characteristics.	30 reciters(19 males and 11 females).	PSD & MFCC	LDA & QDA	Used multiple features extraction and classification methods	Not determined the used dataset
[59]	recognize, identify, and highlight discrepancies between two specific types of Madd rules: the greater connective prolongation and the exchange prolongation rules. This system focuses on verses that contain both rules and aims to point out the mismatches and differences between the rules for Hafss and Warsh recitation styles.	Reciter's database selected from Internet (60 data samples).	MFCC	HMM	-	few data samples
[10]	A system that determines which tajweed rule is used in a specific audio recording of a Quranic recitation (8 tajweed rules).	3,071 audio files collected from ten different expert reciters (5 males and 5 females). Each file contains a recording of one of the 8 rules considered (in either the correct or the incorrect usage).	Traditional (MFCC, LPC, WPD, HMM-SPL) & Non-traditional (CBDN)	KNN, SVM, ANN, RF, multiclass classifier, bagging.	Use of multiple feature extraction algorithms.	-
[57]	A system for distinguishing, recognizing, and correcting the pronunciation of tajweed rules for Warsh narration type).	15 speech simples. 5 verses recited by 3 Warsh .	MFCC	HMM	-	Few data samples
[58]	A system for recognizing, identifying and pointing out the mismatch of the iqlab rules for the verses containing the rules.	6 verses recited by 4 reciters with Qira'at of Warsh. Hence. The total is 24 of speech simples.	MFCC	HMM	-	Few data samples
[14]	differentiate between short and long vowels in Arabic. This distinction is crucial as it plays a significant role in altering the meaning of words.	MFCCs, deltas coefficients, deltas-deltas coefficients and the cepstral pseudoenergy	HMM	have a good accuracy	-	
[37]	A model to identify errors in the Quranic audio files and subsequently distinguish incorrect recitation from the correct recitation.	Ten of expert Qari each of them recite surat Al-Nnass ten times correct and ten times with mistakes.	MFCC	HMM and DTW	-	Covers only one sura
[6]	A system implemented using ASR technique.	Alnass, 20 utterances recited by only a one speaker with and without errors in recitation.	MFCC	ANN	-	Covers only one sura

technique. This setup achieved a WER ranging from 0.27% to 6.31% and a sentence error rate (SER) ranging from 0.4% to 17.39%.

The researchers of [3] used MFCC for the purpose of feature extraction. The researchers adjusted these features using the minimal phone error (MPE) as a discriminative model. The researchers utilized the deep neural network (DNN) model to construct the acoustic model. Here, they introduce an n-gram LM. The dataset utilized for training and assessing the proposed model comprises 10 hours of.wav recitations conducted by 60 reciters. The experimental results demonstrated that the proposed DNN model attained a remarkably low CER of 4.09% and a WER of 8.46%.

Table V summarizes the previous studies of traditional-based speech recognition in the HQSR field.

C. Other HQSR Studies

This section summarizes papers that follow other HQSR techniques, such as using the Google Speech API, Genetic Algorithm (GA), and MFCC.

The authors in [56] used MFCC to detect and recognize sounds for simple IDHAR tajweed without providing any testing results for this study.

Researchers proposed a solution in [22] to facilitate the memorization and learning of the Holy Quran. They employed the Fisher-Yates Shuffle algorithm to randomize the letters of the Quran, aiding in the memorization of verses. In addition, they employed the Jaro-Winkler algorithm for text matching and utilized the Google Speech API for speech recognition. The study focused on data from Juz 30. The achieved accuracy was approximately 91%, with an average matching time of 1.9 ms. However, the study revealed that it was still not possible to distinguish certain Arabic letters with similar pronunciations in Quranic verses in detail.

In [8], the authors produce a new speech segmentation algorithm for the Arabic language. Developing robust algorithms to accurately segment speech signals into fundamental units, rather than just frames, is a crucial preprocessing step in speech recognition systems. They focus on the precise segmentation of Quran recitation using multiple features (entropy, crossings, zero, and energy) and a GA-based optimization scheme. The results of the testing demonstrate a significant enhancement in segmentation performance, with an approximate 20% improvement compared to conventional segmentation techniques based on a single feature.

In [5], the authors implement an Android-based application called TeBook and provide a method for the assessment of the Holy Quran's recitation without the involvement of a third party by taking advantage of the use of speech recognition and an online Holy Quran search engine. There is no testing result present for this study. The limitation of this application is its reliance on multiple online services, which renders it unusable if the services are down.

The authors in [35] suggested a brand-new method called Samee'a to make it easier to memorize any form of literature, including speeches, poetry, and the entire Holy Qur'an. Samee's system utilizes the Jaro Winkler Distance technique to calculate the degree of similarity between the original and transformed texts, and employs the Google Cloud Speech Recognition API to translate Arabic speech to text. Seventy gathered files, ranging in length from twelve to four hundred words, together with a few chapters from the Holy Qur'an, were used to test the system. For the 70 files, the average similarity was 83.33%, while for the chosen chapters of the Holy Qur'an, it was 69%. Preprocessing operations on the text files and the Holy Qur'an improved these results to 91.33% and 95.66%, respectively.

In [48], the authors focus on the digital transformation of Quranic voice signals and the identification of Tajweed-based recitation faults in Harakaat as the primary research objective. They wanted to look into how to process speech using Quranic Recitation Speech Signals (QRSS) in the best digital format possible, using Al-Quran syllables and a design for feature extraction. The objective was to identify similarities or differences in recitation (based on Al-Quran syllables) between experts and students. We employ the DTW approach as a Short Time Frequency Transform (STFT) to quantify the Harakaat of QRSS syllable features. The research presents a method that utilizes human-guidance threshold classification to assess Harakaat, focusing on the syllables of the Qur'an. The categorization performance achieved for Harakaat exceeds 80% in both the training and testing phases.

Table VI summarizes the previously discussed studies of the other techniques used in the HQSR field.

D. HQSR Taxonomy

This section classifies the previously mentioned works of HQSR based on feature extraction methods and classification techniques. It is worth noting that most work done uses MFCCs as feature extraction techniques and the HMM as a classifier (e.g., [37], [14], [58], [57], [59], [45]). Fig. 4 illustrates the techniques of feature extraction and classification used in current HQSR research.

Researchers improved an Arabic recognizer by incorporating a HSMM instead of the traditional HMM [34]. Another approach, mentioned in [6], replaced HMM with ANN. Similarly, Nahar et al. [38] opted for SVM instead of HMM, and in [46] they use CNNs. Furthermore, researchers also explored various feature extraction techniques. For instance, [11] utilized FB for feature extraction and SVM for classification. [33] also used different methods, such as Formant Analysis, Power Spectral Density (PSD), and MFCC, along with LDA and QDA for sorting.

In [10], two categories of feature extraction techniques were employed: traditional and non-traditional. The traditional approach involved the utilization of MFCC, LPC, multi-signal WPD, and HMM-SPL. As for the non-traditional type, they use CDBN. They use K-Nearest Neighbors (KNN), SVM, ANN, Random Forest (RF), multiclass classifiers, and bagging for classification. [50] used MFCC for feature extraction, BLSTM as one of the deep learning topologies, and combined it with HMM as a hybrid system. The entire speech recognition system was built using the Kaldi toolkit [43], starting with feature extraction, acoustic modeling, and model testing. [49] used the deep learning approach in the KALDI toolkit to design, develop, and evaluate an ASR engine for the Holy

TABLE V. SUMMARY OF HQSR TRADITIONAL-BASED SPEECH RECOGNITION

Ref#	Main Idea	Dataset	FE	AM	LM	Pros	Cons
[49]	create a speech recognition engine that is independent of speaker and capable of handling continuous speech. Additionally, a written corpus that accurately represents the script of The Holy Quran is developed. To aid in the recognition process, a phonetic dictionary for The Holy Quran recitations is also be constructed.	32 recitations for Sūrat Taha according to Hafs from A'asim narration.	MFCC	KALDI toolkit to train the acoustic model (traditional and DNN approaches)	n-gram	Best research of current HQSR researches	Used only one sura
[50]	The acoustic model for Quran speech recognition was trained using a deep learning approach. In addition, the model was built to analyze the effect of Quran recitation styles (Maqam) on speech recognition.	The dataset is from everyayah.com [1].	MFCC	Hybrid HMM-BLSTM	3-grams	First work that used BLSTM in HQSR	no preprocessing method to eliminate noise and echo and not specified the used dataset
[21]	Used CMU Sphinx which is a robust framework for speaker-independent continuous speech recognition to train accurate acoustic models.	49 chapters were used: From chapter 067 to chapter 114 in addition to the chapter 001 and the supplication that is recited before the Holy Quran (Isti'adah).	CMU Sphinx framework			Get WER around 15% of trained acoustic model.	-
[34]	Presented the results of an enhanced Arabic recognizer by implementing an HSMM model instead of the standard one utilized in the baseline recognizer.	Arabic database utilized consists of 5935 waveform files for 10 reciters.	MFCC	HSMM	flat LM	Get enhancement by around 1.5% in the accuracy	-
[3]	Suggested the traditional method to recognizing Qur'an verses using a dataset of Qur'an verses.	A total duration of 10 hours of MP3 recordings containing recitations of Qur'an verses by 60 reciters.	MFCC	DNN	n-gram	Well-structured and clear article.	-

TABLE VI. SUMMARY OF OTHER HQSR STUDIES

Ref#	Main Idea	Dataset	Used Algorithms	Pros	Cons
[5]	Allow learners to learn how to memorize without the constraints of being in a fixed place and outside the classroom.	Everyayah.com Recitation Audio, Surah.my Translation	Alfanous JOS2 API, Android Speech Recognition	-	Relied heavily on online services and not specified the used dataset
[8]	A novel speech segmentation algorithm for Arabic language with a focus on the accurate segmentation of Quran recitation. Starting with a set of initial segmentations, three basic speech features: zero crossings, entropy, and energy are used.	They used the comprehensive KACST dataset with manually labelled Quran syllable structures.	feature fusion and Genetic Algoritms.	First segmentation on Quran recitation.	-
[22]	A solution to memorize and learn the Holy Quran easily.	juz 30	Fisher-Yates Shuffle Jaro-Winkler	-	Relies on Google Speech API which not trained on Quran verses
[56]	Emphasize idhar, which had a distinct and unambiguous pronunciation. The chosen hijaiyah letters comprised six possibilities, with only nun sukun and tanwin, making it effortless to identify them.	-	MFCC, FFT	-	Not specifying dataset, no testing result, and the poor organization of the content.
[35]	This article introduces a novel system called Samee'a, which aims to enhance the process of memorizing various types of texts, including poems, speeches, and the Holy Qur'an.	The system completed testing utilizing a dataset of 70 files, with word counts ranging from 12 to 400, including selected chapters from the Holy Qur'an.	Google Cloud Speech Recognition API and Jaro Winkler Distance algorithm	A comprehensive and informative paper	-
[48]	This study focuses on the digital transformation of Quranic voice signals and the identification of Tajweed-based recitation faults of Harakaat as its main research objective.	-	Dynamic Time Warping (DTW)	-	Not specifying dataset

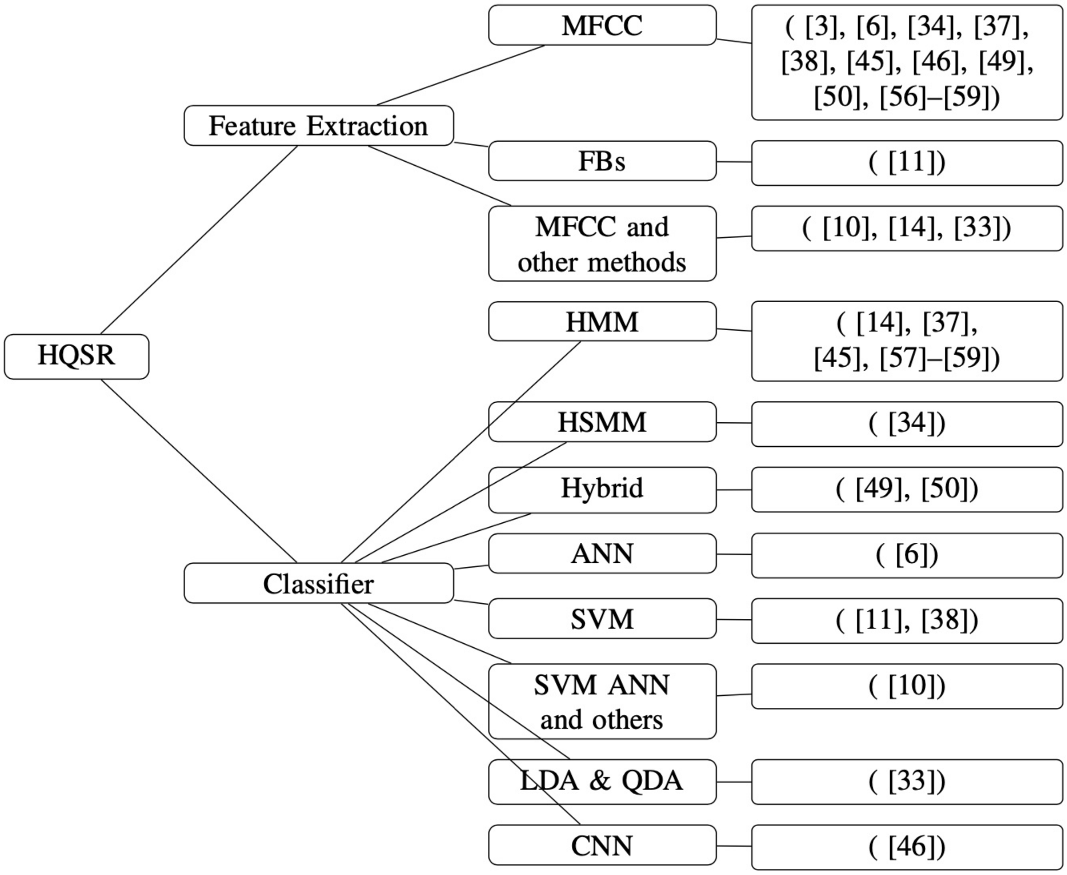

Fig. 4. HQSR Taxonomy of Used Technique

Quran recitations. The best experimental setup was achieved using TDNN with sub-sampling technique.

In [21], the CMU Sphinx trainer [53] was employed to train the acoustic model specifically for the Holy Quran. In a similar vein, a study by [22] utilized the Jaro-Winkler algorithm for text matching and relied on the Google Speech API to establish a framework for speech recognition.

The solution of [5] uses Android speech recognition and depends heavily on third-party online services. In [8], they developed a robust hybrid speech segmentation system based on multiple features (entropy, zero crossings, and energy) and a GA-based optimization scheme to obtain accurate segment units specially adapted for Quran recitation.

IV. Discussion and Future Research Directions

Recent research in the field of HQSR has suggested numerous works. We provide in Table VII a comparative analysis of current research works based on the dataset characteristics and the suggested methodology:

- Dataset characteristics:
  1) #verses: This refers to the total number of verses used in the study: L (number of verses between 1-100), M (number of verses between 101-200), H (number of verses greater than 200), and N (number of verses not determined in the paper).
  2) #sura: This refers to the total number of suras used in the study.
  3) #reciters: This shows the number of reciters who participate in this study.
- Proposed methodology:
  1) DL-based: This criteria shows if the study used deep learning in any stage of the study.
  2) LM: This shows if the study used a language model in their solution or not.
  3) AM: This shows if the study used an acoustic model in their solution or not.
  4) Reciter Independent: This shows if the output model of this study is reciter independent or not.
  5) Speaker Adaptation: This shows if this study used any techniques of speaker adaptation or not.

As we can see in Table VII, there is a significant gap in the current work of HQSR. First, most of the works in HQSR follow template-based speech recognition, which is an old style of speech recognition. This style extracts features of raw audio and feeds these features into a classifier to classify and match with stored templates without using acoustic, lexical (pronunciation), or language models [58]. Examples of these works are ([45], [38], [11], [33], [59], [57], [58], [14], [6], and [37]) as shown in Table VII. Few works suggest the use of deep learning. However, they still rely on an old-style design. For instance, [10] used a deep learning architecture with the old style (i.e., it didn't use acoustic, lexical, and language models but deep learning in the feature extraction phase only). In addition, the authors in [5], [22] employed the Google Speech API for their solution. However, this approach had limitations as the API was unable to accurately differentiate between the seven letters that share similar pronunciations in the verses of the Quran.

Second, only a few studies follow traditional speech recognition, as explained in Fig. 2, either with a deep learning architecture like [49], [50], [3] or without, such as [34], [21]. It is worth noting that more work investigating deep learning architecture should be conducted to improve the accuracy of Arabic speech recognition in general and the Holy Quran in particular.

Third, we can observe in Table VII that the used data set is too small for most of the work (number of verses, suras, and reciters). while an extensive dataset helps produce a robust and generalized speech recognition system.

Finally, no research on HQSR used end-to-end deep learning architecture, while this architecture shows outstanding results with Arabic and non-Arabic languages, as previously discussed in Section II-B (see Table III, which presents a comparison between some end-to-end based speech recognition architecture and traditional techniques in Arabic and non-Arabic languages).

To sum up, many challenges still need to be considered in future work. Indeed, in recognition of speech, recognizing individual words is easy, but the challenge is recognizing continuous speech [7]. Multiple conditions, including speaker dependency, vocabulary size, and noisy environments, can affect the performance of speech recognition systems. Recognition performance increases with limited vocabulary and speaker-dependent conditions while using broad vocabulary and speaker-independent scenarios; performance can decrease significantly [7]. Moreover, Arabic is a morphologically complex language that contains a high degree of affixation and derivation, resulting in a massive increase in word forms [31]. Furthermore, speech recognition of the Holy Quran has additional difficulties compared with written and spoken languages for the following reasons:

- Lack of a comprehensive dataset that contains recitations of women, children, and native and non-native Arabic speakers with both the correct and incorrect recitation of the Holy Quran.
- Mistakes are not acceptable when reading the Quran because an error in reciting only one letter may change the meaning.
- The diversity of narrations in reading the Qur'an makes it difficult for the model to recognize different narrations.
- The diversity of *Magam* in Quran Recitation, such as (*bayat*, *Ajam*, *Nahawand*, *Hijaz*, *Rost*, *Sika*, etc.), adds more difficulty for the model when recognizing the recitation.
- The length of prolongation (*Madd*) varies when reciting the Quran. In Hafs An Asim narration, reciters can recite some types of the madd with 2, 4, or 5 *Harakat*.
- Recitation of the Holy Quran must follow the rules of "tajweed" and correctly pronounce Makhraj (point of articulations) and the Sifaat (characteristics) of each alphabet.

TABLE VII. COMPARING HQSR SOLUTIONS

Ref#	Dataset			Methodology				
	#verses	#sura	#reciters	DL-based	LM	AM	Reciter Independent	Speaker Adaptation
[45]	L	1	1					
[49]	M	1	32	✓	✓	✓	✓	✓
[38]	H							
[11]	H							
[33]	N		30					
[5]	N							
[50]	N		13	✓	✓	✓		✓
[22]	H							
[59]	L							
[10]	H		10	✓				
[57]	L							
[58]	L							
[56]	N							
[21]	H	49	39		✓	✓	✓	
[34]	H		10		✓	✓		
[14]	H	3	4					
[37]	L		10					
[6]	L	1	1					
[46]	H		7	✓				
[3]	L		60	✓	✓	✓		

Note: *L (number of verses between 1-100), M (number of verse between 101-200), H (number of verses greater than 200), and N (Number of verses not determined in the paper).*

V. CONCLUSION

This paper surveys Holy Quran Speech Recognition (HQSR) works. It summarizes some studies of speech recognition in written and spoken languages and the most recent work in the HQSR field. It provides a general taxonomy of speech recognition and a specific one dedicated to HQSR studies that illustrates the techniques of feature extraction and classification used in current HQSR research. We compared the current solutions and clarified the limitations of the current studies. The main challenges of the HQSR field are the lack of a comprehensive dataset, minimizing mistakes that are not acceptable when reading the Quran, diversity of narrations, diversity of Magam in Quran recitation, and diversity of prolongation (Madd) length when reciting the Quran. The field of HQSR needs a lot of work to improve the current speech recognition models of the Holy Quran by using better techniques that already show good results with written and spoken languages but haven't been used with HQSR yet.

ACKNOWLEDGMENT

This research work was funded by Institutional Fund Projects under grant no. (IFPIP-1340-612-1443). The authors gratefully acknowledge the technical and financial support provided by the Ministry of Education and King Abdulaziz University, Deanship of Scientific Research, Jeddah, Saudi Arabia.

References

[1] Every ayah, http://www.everyayah.com/, 2022.

[2] Muslim population by country 2021, https://worldpopulationreview.com/country-rankings/muslim-population-by-country, 2021.

[3] Alsayadi Hamzah A and Hadwan Mohammed. Automatic speech recognition for qur'an verses using traditional technique. *Journal of Artificial Intelligence and Metaheuristics (JAIM)*, 2022.

[4] Abdelaziz A Abdelhamid, Hamzah A Alsayadi, Islam Hegazy, and Zaki T Fayed. End-to-end arabic speech recognition: A review. 2020.

[5] Mohd Hafiz Bin Abdullah, Zalilah Abd Aziz, Rose Hafsah Abd Rauf, Noratikah Shamsudin, and Rosmah Abd Latiff. Tebook a mobile holy quran memorization tool. In *2019 2nd International Conference on Computer Applications & Information Security (ICCAIS)*, pages 1–6. IEEE, 2019.

[6] Bushra Abro, Asma Batool Naqvi, and Ayyaz Hussain. Qur'an recognition for the purpose of memorisation using speech recognition technique. In *2012 15th International Multitopic Conference (INMIC)*, pages 30–34. IEEE, 2012.

[7] Ahmed Hamdi Abo Absa. *Self-Learning Techniques for Arabic Speech Segmentation and Recognition*. Thesis, 2018.

[8] Ahmed Hamdi Abo Absa, Mohamed Deriche, Moustafa Elshafei-Ahmed, Yahya Mohamed Elhadj, and Biing-Hwang Juang. A hybrid unsupervised segmentation algorithm for arabic speech using feature fusion and a genetic algorithm (july 2018). *IEEE Access*, 6:43157–43169, 2018.

[9] Abdelrahman Ahmed, Yasser Hifny, Khaled Shaalan, and Sergio Toral. End-to-end lexicon free arabic speech recognition using recurrent neural networks. *Computational Linguistics, Speech And Image Processing For Arabic Language*, pages 231–248, 2019.

[10] Mahmoud Al-Ayyoub, Nour Alhuda Damer, and Ismail Hmeidi. Using deep learning for automatically determining correct application of basic quranic recitation rules. *Int. Arab J. Inf. Technol.*, 15(3A):620–625, 2018.

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

[12] Abdulmalik A Alasadi, TH Aldhayni, Ratnadeep R Deshmukh, Ahmed H Alahmadi, and Ali Saleh Alshebami. Efficient feature extraction algorithms to develop an arabic speech recognition system. *Engineering, Technology & Applied Science Research*, 10(2):5547–5553, 2020.

[13] Hanan Aldarmaki, Asad Ullah, and Nazar Zaki. Unsupervised automatic speech recognition: A review. *arXiv preprint arXiv:2106.04897*, 2021.

[14] Yousef A Alotaibi, Mohammed Sidi Yakoub, Ali Meftah, and Sid-Ahmed Selouani. Duration modeling in automatic recited speech recognition. In *2016 39th International Conference on Telecommunications and Signal Processing (TSP)*, pages 323–326. IEEE, 2016.

[15] Fatimah Alqadheeb, Amna Asif, and Hafiz Farooq Ahmad. Correct pronunciation detection for classical arabic phonemes using deep learning. In *2021 International Conference of Women in Data Science at Taif University (WiDSTaif)*, pages 1–6. IEEE, 2021.

[16] Hamzah A Alsayadi, Abdelaziz A Abdelhamid, Islam Hegazy, and Zaki T Fayed. Arabic speech recognition using end-to-end deep learning. *IET Signal Processing*, 2021.

[17] Dario Amodei, Sundaram Ananthanarayanan, Rishita Anubhai, Jingliang Bai, Eric Battenberg, Carl Case, Jared Casper, Bryan Catanzaro, Qiang Cheng, and Guoliang Chen. Deep speech 2: End-to-end speech recognition in english and mandarin. In *International conference on machine learning*, pages 173–182. PMLR, 2016.

[18] Chung-Cheng Chiu, Tara N Sainath, Yonghui Wu, Rohit Prabhavalkar, Patrick Nguyen, Zhifeng Chen, Anjuli Kannan, Ron J Weiss, Kanishka Rao, and Ekaterina Gonina. State-of-the-art speech recognition with sequence-to-sequence models. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4774–4778. IEEE, 2018.

[19] Ronan Collobert, Awni Hannun, and Gabriel Synnaeve. A fully differentiable beam search decoder. In *International Conference on Machine Learning*, pages 1341–1350. PMLR, 2019.

[20] Ronan Collobert, Christian Puhrsch, and Gabriel Synnaeve. Wav2letter: an end-to-end convnet-based speech recognition system. *arXiv preprint arXiv:1609.03193*, 2016.

[21] Mohamed Yassine El Amrani, MM Hafizur Rahman, Mohamed Ridza Wahiddin, and Asadullah Shah. Towards an accurate speaker-independent holy quran acoustic model. In *2017 4th IEEE International Conference on Engineering Technologies and Applied Sciences (ICETAS)*, pages 1–4. IEEE, 2017.

[22] YA Gerhana, AR Atmadja, DS Maylawati, A Rahman, K Nufus, H Qodim, and MA Ramdhani. Computer speech recognition to text for recite holy quran. In *IOP Conference Series: Materials Science and Engineering*, volume 434, page 012044. IOP Publishing, 2018.

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

[24] Imane Guellil, Houda Saâdane, Faical Azouaou, Billel Gueni, and Damien Nouvel. Arabic natural language processing: An overview. *Journal of King Saud University-Computer and Information Sciences*, 33(5):497–507, 2021.

[25] Hossein Hadian, Hossein Sameti, Daniel Povey, and Sanjeev Khudanpur. End-to-end speech recognition using lattice-free mmi. In *Interspeech*, pages 12–16, 2018.

[26] Awni Hannun, Ann Lee, Qiantong Xu, and Ronan Collobert. Sequence-to-sequence speech recognition with time-depth separable convolutions. *arXiv preprint arXiv:1904.02619*, 2019.

[27] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[28] Xiaodong He and Li Deng. Discriminative learning for speech recognition: theory and practice. *Synthesis Lectures on Speech and Audio Processing*, 4(1):1–112, 2008.

[29] Ahmed Ali Hifny, Shammur Absar Chowdhury, Amir Hussein, and Yasser. Arabic code-switching speech recognition using monolingual data. *Proc. Interspeech 2021*, pages 3475–3479, 2021.

[30] Geoffrey Hinton, Li Deng, Dong Yu, George E Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, and Tara N Sainath. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal processing magazine*, 29(6):82–97, 2012.

[31] Amir Hussein, Shinji Watanabe, and Ahmed Ali. Arabic speech recognition by end-to-end, modular systems and human. *arXiv preprint arXiv:2101.08454*, 2021.

[32] Biing Hwang Juang and Laurence R Rabiner. Hidden markov models for speech recognition. *Technometrics*, 33(3):251–272, 1991.

[33] Safiah Khairuddin, Salmiah Ahmad, Abdul Halim Embong, Nik Nur Wahidah Nik Hashim, and Surul Shahbuddin Hassan. Features identification and classification of alphabet (ro) in leaning (al-inhiraf) and repetition (al-takrir) characteristics. In *2019 IEEE International Conference on Automatic Control and Intelligent Systems (I2CACIS)*, pages 295–299. IEEE, 2019.

[34] Mohamed OM Khelifa, Mostafa Belkasmi, Yousfi Abdellah, and Yahya OM ElHadj. An accurate hsmm-based system for arabic phonemes recognition. In *2017 Ninth International Conference on Advanced Computational Intelligence (ICACI)*, pages 211–216. IEEE, 2017.

[35] Souad Larabi-Marie-Sainte, Betool S. Alnamlah, Norah F. Alkassim, and Sara Y. Alshathry. A new framework for arabic recitation using speech recognition and the jaro winkler algorithm. *Kuwait Journal of Science*, 49, 2022.

[36] Lina Marlina, Cipto Wardoyo, WS Mada Sanjaya, Dyah Anggraeni, Sinta Fatmala Dewi, Akhmad Roziqin, and Sri Maryanti. Makhraj recognition of hijaiyah letter for children based on mel-frequency cepstrum coefficients (mfcc) and support vector machines (svm) method. In *2018 International Conference on Information and Communications Technology (ICOIACT)*, pages 935–940. IEEE, 2018.

[37] Ammar Mohammed, Mohd Shahrizal Sunar, and Md Sah Hj Salam. Quranic verses verification using speech recognition techniques. *Jurnal Teknologi*, 73(2), 2015.

[38] Khalid MO Nahar, M Ra'ed, A Moy'awiah, and M Malek. An efficient holy quran recitation recognizer based on svm learning model. *Jordanian Journal of Computers and Information Technology (JJCIT)*, 6(04), 2020.

[39] Maryam Najafian, Wei-Ning Hsu, Ahmed Ali, and James Glass. Automatic speech recognition of arabic multi-genre broadcast media. In *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 353–359. IEEE, 2017.

[40] Maryam Najafian, Sameer Khurana, Suwon Shan, Ahmed Ali, and James Glass. Exploiting convolutional neural networks for phonotactic based dialect identification. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5174–5178. IEEE, 2018.

[41] Lv Ping. English speech recognition method based on hmm technology. In *2021 International Conference on Intelligent Transportation, Big Data & Smart City (ICITBS)*, pages 646–649. IEEE, 2021.

[42] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, and Petr Schwarz. The kaldi speech recognition toolkit. In *IEEE 2011 workshop on automatic speech recognition and understanding*. IEEE Signal Processing Society, 2011.

[43] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, and Petr Schwarz. The kaldi speech recognition toolkit. In *IEEE 2011 workshop on automatic speech recognition and understanding*. IEEE Signal Processing Society, 2011.

[44] Rohit Prabhavalkar, Kanishka Rao, Tara N Sainath, Bo Li, Leif Johnson, and Navdeep Jaitly. A comparison of sequence-to-sequence models for speech recognition. In *Interspeech*, pages 939–943, 2017.

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

[46] Ghassan Samara, Essam Al-Daoud, Nael Swerki, and Dalia Alzu'bi. The recognition of holy qur'an reciters using the mfccs' technique and deep learning. *Advances in Multimedia*, 2023, 2023.

[47] Benjamin Elisha Sawe. Arabic speaking countries, Jul 2018.

[48] Noraimi Shafie, Azizul Azizan, Mohamad Zulkefli Adam, Hafiza Abas, Yusnaidi Md Yusof, and Nor Azurati Ahmad. Dynamic time warping features extraction design for quranic syllable-based harakaat assessment. *International Journal of Advanced Computer Science and Applications*, 13, 2022.

[49] Imad K Tantawi, Mohammad AM Abushariah, and Bassam H Hammo. A deep learning approach for automatic speech recognition of the holy qur'an recitations. *International Journal of Speech Technology*, pages 1–16, 2021.

[50] Faza Thirafi and Dessi Puji Lestari. Hybrid hmm-blstm-based acoustic modeling for automatic speech recognition on quran recitation. In *2018 International Conference on Asian Language Processing (IALP)*, pages 203–208. IEEE, 2018.

[51] Pahini A Trivedi. Introduction to various algorithms of speech recognition: Hidden markov model, dynamic time warping and artificial neural networks. *International Journal of Engineering Development and Research*, 2(4):3590–3596, 2014.

[52] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, ?ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

[53] Willie Walker, Paul Lamere, Philip Kwok, Bhiksha Raj, Rita Singh, Evandro Gouvea, Peter Wolf, and Joe Woelfel. Sphinx-4: A flexible open source framework for speech recognition. 2004.

[54] Song Wang and Guanyu Li. Overview of end-to-end speech recognition. In *Journal of Physics: Conference Series*, volume 1187, page 052068. IOP Publishing, 2019.

[55] Shinji Watanabe, Takaaki Hori, Shigeki Karita, Tomoki Hayashi, Jiro Nishitoba, Yuya Unno, Nelson Enrique Yalta Soplin, Jahn Heymann, Matthew Wiesner, and Nanxin Chen. Espnet: End-to-end speech processing toolkit. *arXiv preprint arXiv:1804.00015*, 2018.

[56] Efy Yosrita and Abdul Haris. Identify the accuracy of the recitation of al-quran reading verses with the science of tajwid with mel-frequency ceptral coefficients method. In *2017 International Symposium on Electronics and Smart Devices (ISESD)*, pages 179–183. IEEE, 2017.

[57] Bilal Yousfi and Akram M Zeki. Holy qur'an speech recognition system imaalah checking rule for warsh recitation. In *2017 IEEE 13th international colloquium on signal processing & its applications (CSPA)*, pages 258–263. IEEE, 2017.

[58] Bilal Yousfi, Akram M Zeki, and Aminah Haji. Isolated iqlab checking rules based on speech recognition system. In *2017 8th International Conference on Information Technology (ICIT)*, pages 619–624. IEEE, 2017.

[59] Bilal Yousfi, Akram M Zeki, and Aminah Haji. Holy qur'an speech recognition system distinguishing the type of prolongation. *Sukkur IBA Journal of Computing and Mathematical Sciences*, 2(1):36–43, 2018.

[60] Naima Zerari, Samir Abdelhamid, Hassen Bouzgou, and Christian Raymond. Bidirectional deep architecture for arabic speech recognition. *Open Computer Science*, 9(1):92–102, 2019.

[61] Yaxin Zhang, Mike Alder, and Roberto Togneri. Using gaussian mixture modeling in speech recognition. In *Proceedings of ICASSP'94. IEEE International Conference on Acoustics, Speech and Signal Processing*, volume 1, pages I/613–I/616 vol. 1. IEEE, 1994.