IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 23 November 2023, accepted 13 December 2023, date of publication 22 December 2023, date of current version 3 January 2024.

Digital Object Identifier 10.1109/ACCESS.2023.3345843

RESEARCH ARTICLE

Improving Automatic Forced Alignment for Phoneme Segmentation in Quranic Recitation

AMMAR MOHAMMED ALI ALQADASI[1,2,3], AKRAM M. ZEKI[2], (Senior Member, IEEE), MOHD SHAHRIZAL SUNAR[3,4], (Member, IEEE), MD. SAH BIN HJ SALAM[3], RAWAD ABDULGHAFOR[2,5], AND NASHWAN ABDO KHALED[6]

[1]Al-Saeed Faculty of Engineering and Information Technology, Taiz University, Taiz, Yemen
[2]Faculty of Information and Communication Technology, International Islamic University Malaysia, Kuala Lumpur 53100, Malaysia
[3]Faculty of Computing, Universiti Teknologi Malaysia, Johor Bahru 81310, Malaysia
[4]Media and Game Innovation Centre of Excellence, Institute of Human Centered Engineering, Universiti Teknologi Malaysia, Johor Bahru 81310, Malaysia
[5]Faculty of Computer Studies (FCS), Arab Open University Oman, Muscat 130, Oman
[6]AbdulHamid A. AbuSulayman Kulliyyah of Islamic Revealed Knowledge and Human Sciences, International Islamic University Malaysia, Kuala Lumpur 53100, Malaysia

Corresponding author: Ammar Mohammed Ali Alqadasi (ammaralqadasi2@gmail.com)

ABSTRACT Segmentation plays a crucial role in speech processing applications, where high accuracy is essential. The quest for improved accuracy in automatic segmentation, particularly in the context of the Arabic language, has garnered substantial attention. However, the differences between Qur'an recitation and normal Arabic speech, especially with regard to intonation rules affecting the lengthening of long vowels, pose challenges in segmentation especially for Qur'an recitation. This research endeavors to address these challenges by delving into the domain of automatic segmentation for Qur'an recitation recognition. The proposed scheme employs a hidden Markov models (HMMs) forced alignment algorithm. To enhance the precision of segmentation, several refinements have been introduced, with a primary emphasis on the phonetic model of the Qur'an and Tajweed, particularly the intricate rules governing elongation. These enhancements encompass the adaptation of an acoustic model tailored for Qur'anic recitation as preprocessing and culminate in the development of an algorithm aimed at refining forced alignment based on the phonetic nuances of the Qur'an. These enhancements are seamlessly integrated as post-processing components for the classic HMM-based forced alignment. The research utilizes a comprehensive database featuring recordings from 100 renowned Qur'an reciters, encompassing the recitation of 21 Qur'anic verses (Ayat). Additionally, 30 reciters were asked to record the same verses, incorporating various recitation speed patterns. To facilitate the evaluation process, a Random sample of the Qur'anic database was manually segmented, comprised 21 Ayats, totaling 19,800 words, with 89 unique words (14 verses x 3 recitation levels: fast, slow and normal x 6 readers). The outcomes of this study manifest notable advancements in the alignment of long vowels within Qur'an recitation, all while maintaining the precise alignment of vowels and consonants. Objective comparisons between the proposed automatic methods and manual segmentation were conducted to ascertain the superior approach. The findings affirm that the classic forced alignment method produces satisfactory outcomes when employed on verses lacking long vowels. However, its performance diminishes when confronted with verses containing long vowels. Therefore, the test samples were categorized into three groups based on the presence of long vowels, resulting in a Correct Classification Rate (CCR) that ranged from 6% to 57%, contingent on whether the verse includes long vowels or not. The average CCR across all test samples was 23%. In contrast, the proposed algorithm significantly enhances audio segmentation. It achieved CCR values ranging from 16% to 70% within the same database categories, with an average CCR of 45% across all test samples. This marks a notable advancement of 22% in segmented speech accuracy, particularly within a 30 ms tolerance, for verses containing long vowels.

The associate editor coordinating the review of this manuscript and approving it for publication was Ghulam Muhammad.

INDEX TERMS Phoneme alignment, forced alignment, phoneme segmentation, Arabic phoneme segmentation, phoneme duration, phoneme recognition, recitation recognition, Tajweed recognition.

I. INTRODUCTION

In the realm of speech processing, the fundamental task of speech segmentation holds a pivotal role. This process involves dissecting continuous or connected speech into discrete, meaningful units [1], identifying the natural boundaries within spoken language [2], [3]. These segmented units can encompass phonemes [4], syllables [5], words [6], [7], or even sub-words and sentences [8].

Segmentation, an indispensable step in various speech processing applications such as speech recognition and synthesis [9], necessitates a high degree of accuracy to ensure the proper functioning of these systems. Additionally, segmentation plays a vital role in detecting the precise start and end points of speech, optimizing memory utilization, and reducing computational complexity, especially in large vocabulary systems [3]. Furthermore, it holds significance in the creation of datasets and file labeling, a crucial phase in the development of speech processing systems, used for training and testing.

Speech segmentation transcends its importance within the realm of speech processing; it extends its relevance to addressing challenges in Natural Language Processing (NLP) as well. It serves a multifaceted purpose by preparing datasets for training, expanding databases automatically, labeling public databases, mitigating background noise, and pinpointing the necessary samples by delineating their starting and ending points.

However, the task of speech segmentation is not without its challenges. Unlike written text, spoken utterances lack the inherent acoustic gaps between word utterances [2], [3] Moreover, in continuous speech, phonemes tend to merge, and this challenge is particularly pronounced with consonants due to their brief duration [10]. Additionally, variations in speakers and languages introduce further complexities, as different intonations, rhythms, pitch changes, and distinct sounds emerge within language [11], [12].

Amidst this intricate landscape of speech segmentation, the Arabic language emerges as a distinct player. Arabic is renowned for its historical continuity, remaining remarkably consistent over centuries, primarily due to its role as the language of the Qur'an, the holy book of Islam, which is recited and written in classical Arabic. Muslims, as a religious practice, teach their children to read the Qur'an in classical Arabic, further ensuring its continuity [13]. However, it's essential to differentiate between Classical Arabic (the language of the Qur'an), Modern Standard Arabic (MSA), and Dialect Arabic (DA) [14].

The Arabic language encompasses 44 phonemes in all three forms, which include consonants, short vowels (ضمة و فتحة وكسرة الحركات), and long vowels (حروف العلة /حروف المد). Notably, long vowels, particularly in the context of Tajweed rules, introduce variations in Arabic pronunciation [16]. Tajweed rules dictate the pronunciation duration, especially for Medd types (المد) which express the time taken by the reciter (قارئ) to articulate the vowel. While normal Arabic speech divides vowels into short (1 Harakah) and long (2 Harakah), in Qur'anic recitation, this expands to include durations of 4 Harakah, 5 Harakah, and 6 Harakah, known as Medd in Tajweed rules [15], [16].

These unique aspects of Qur'anic recitation present a distinct challenge in phoneme segmentation, amplifying the limitations of the forced alignment algorithm when applied in this context. Nevertheless, the well-defined rules governing phoneme duration, including Medd types, offer avenues to enhance phoneme segmentation accuracy. In light of the deep connection between pronunciation and the Arabic language's phonetics and its relationship with the rules of Tajweed, the development of a system that regulates sound intricacies greatly aids in sound control and serves the rules of Tajweed themselves.

In this study, our primary contribution revolves around the creation of a specialized speech segmentation algorithm rooted in the acoustic characteristics of Qur'anic recitation. We have integrated an optimization framework into the Hidden Markov Model Toolkit (HTK) forced alignment algorithm, meticulously tailored to ensure precision in phoneme segmentation for Qur'anic recitation. The process begins with an initial set of segmentations generated using the HTK forced alignment algorithm.

This paper focuses on the specific context of Qur'anic recitation, where the duration of vocalic phonemes adheres to distinct traditions compared to normal, conversational Arabic. To address this unique recitation style, we propose refinements to the foundational segmentation algorithm and provide empirical evidence through experimental results.

The structure of this paper is as follows: Section II provides a comprehensive literature review of speech segmentation techniques. Section III delves into a detailed discussion of our proposed work. In Section IV, we elucidate the performance evaluation of our segmentation methods, accompanied by the presentation of experimental results. Finally, Section V offers a conclusive summary of our findings.

II. LITERATURE REVIEW

The fundamental notion underpinning segmentation involves the partitioning of continuous speech into discrete, smaller entities such as sentences, words, syllables, semi-syllables, letters, or phonemes, contingent upon the intended purpose of the segmentation [5]. This process of breaking down these smaller entities is accomplished by discerning distinct acoustic attributes within adjacent segments. Various endeavors by researchers have been made towards the development of automated segmentation techniques. However, the pinnacle of accuracy in speech segmentation is typically achieved through manual methods. Nevertheless, manual segmentation comes with its drawbacks, including time constraints and high costs, as it necessitates the engagement of skilled human experts, and introduces an element of subjectivity into the segmentation task [17].

Within this section, we provide an overview of the different types of segmentation and delve into a discussion of noteworthy studies pertaining to the segmentation of Qur'anic recitation and the Arabic language.

A. OVERVIEW OF SPEECH SEGMENTATION METHODS

There are two main categories of speech recognition: (i) manual segmentation and (ii) automatic segmentation.

1) MANUAL SEGMENTATION

Manual segmentation and annotation are renowned for their precision in marking speech boundaries and phonetic units, aligning closely with the nuanced aspects of spoken language. Typically, this meticulous task is entrusted to skilled domain experts, well-versed in linguistic subtleties and phonetics.

However, this method is not without its challenges. The process of manual segmentation and annotation can be painstakingly slow and labor-intensive, demanding a substantial amount of time even for modest speech datasets. Furthermore, it necessitates a level of expertise that might not always be readily available, especially for languages with intricate phonological features or specialized contexts like Qur'anic recitation.

An additional concern with manual segmentation is the potential for variability introduced by human judgment. Even among expert annotators, differences in auditory and visual interpretations can lead to inconsistencies in segmentation, impacting the reliability and reproducibility of research findings.

Given these limitations, there is growing interest in automatic speech segmentation methods within the field of speech processing. These automated techniques employ computational approaches and machine learning, to detect and define speech segments and phonetic units. Researchers actively pursue the development of more accurate and efficient automatic segmentation methods, with the goal of achieving the precision of manual segmentation while capitalizing on the speed and scalability benefits that automation offers.

2) AUTOMATIC SEGMENTATION

Automatic speech segmentation can be categorized into two main types, often referred to as (i) Blind segmentation and (ii) Aided segmentation algorithms [3], also known as supervised and unsupervised methods [18].

In blind segmentation, there is a notable absence of prior knowledge about linguistic properties, such as orthography or full phonetic annotations for the speech segments. Instead, these methods solely rely on acoustic features extracted from the signal. Commonly utilized features include Mel Frequency Cepstral Coefficients (MFCCs), Linear prediction (LP-coefficients), or the pure Fast Fourier Transform (FFT) spectrum. These techniques find application in various fields, including speech corpus segmentation and labeling, speech recognition systems, speaker verification systems, and language identification systems.

Numerous studies have employed various blind segmentation methods. For instance, Rahman et al. [19] introduced a method involving the use of blocking black areas to distinguish voiced regions from silence in continuous speech for Bangla sentence segmentation using MATLAB. Kaur and Singh implemented the Short Term Energy of speech method in MATLAB to segment Punjabi speech into syllables [20]. A similar approach was used by Hossain et al. [21] for the automatic classification of Bangla velar phonemes into silence/unvoiced/voiced. Sharma and Singh proposed a three-stage technique involving feature extraction, rule matching, and segmentation using Word Chopper to segment speech into syllables [22]. Tolba et al. [23] presented a technique using the Wavelet Method, devoid of linguistic information, focusing on detecting transitions between consonant and vowel segments in Arabic speech.

In aided segmentation, external linguistic knowledge of the speech is incorporated. This often includes orthographic or phonetic transcription data, which contributes in parallel to the speech or training algorithms. Methods falling into this category encompass techniques using Hidden Markov Models (HMMs) [24], Dynamic Time Warping (DTW) [25], or Artificial Neural Networks (ANNs) [26], [27].

A substantial body of research adopts aided segmentation techniques. For example, Ziółko [25] employed Discrete Wavelet Transform (DWT) analysis using the Wavelet Method to identify the start and end of phonemes. Kamarauskas [28] utilized perceptron and backpropagation artificial neural networks to recognize distinctive phonemes, with backpropagation showing lower recognition error rates but requiring longer training times. Bansal et al. [29] used HMM-based methods for phonetic segmentation and speech analysis at the phonetic level. Dines et al. [30] implemented segmentation as a step in their speech synthesis research, employing HMM training strategies optimized for the task to enhance segmentation reliability. Stolcke et al. [31] demonstrated improvements in automatic phonetic segmentation accuracy using HMM-based models conditioned on phonetic context and duration features.

In general, it is noteworthy that since the 1980s, HMM-based algorithms have played a dominant role in most speech recognition applications due to their remarkable recognition performance and computational efficiency in the field of speech recognition [5]. With Hidden Markov Models (HMMs) being a common choice. However, these methods often rely on manually labeled data, which is scarce for many languages. For example to tackle this challenge in the context of Indian languages, the research presented in [32] introduces an innovative strategy. This approach substitutes Gaussian mixture models (GMMs) within HMM models with deep neural networks (DNNs) and convolutional neural networks (CNNs) to perform speech segmentation in Indian language text-to-speech (TTS) systems.

In conclusion, Hidden Markov models (HMMs) hold significant historical importance within the realm of speech processing. Nevertheless, their practical utility has been hampered by the requirement for manual data labeling and their less-than-ideal performance, notably evident when dealing with intricate languages such as Arabic. In response to these limitations, inventive approaches have emerged to enhance their efficacy. These approaches encompass the substitution of Gaussian mixture models (GMMs) with deep neural networks (DNNs) and convolutional neural networks (CNNs), in addition, the integration of post-processing techniques aimed at bolstering accuracy. These advancements have showcased their utmost value, particularly in linguistic domains where annotated resources are scarce and when confronting languages characterized by complex linguistic features like Arabic. The subsequent section offers a comprehensive overview of recent research endeavors that have leveraged or developed and refined HMMs-based forced alignment methodologies to address these challenges.

B. OVERVIEW OF RECENT ADVANCES IN FORCED ALIGNMENT METHOD

Forced alignment has been a foundational component of speech recognition technology since the 1970s and found its initial applications in acoustics research during the 1990s [33]. However, it continues to captivate the attention of researchers across various domains, including phonetic segmentation, speech segmentation, and speech processing applications like speech synthesis and speech recognition. Researchers have been drawn to forced alignment for diverse reasons, including its adaptation to novel linguistic contexts and environments, enhancement to boost accuracy, and mitigation of inherent limitations and drawbacks [34]. Furthermore, certain researchers have utilized forced alignment as a benchmark for assessing the effectiveness of their phonetic segmentation techniques, comparing their results with those produced by forced alignment [35].

Wohlan et al. [36] emphasizes the critical role of phoneme segmentation in healthcare applications, particularly for diagnosing and monitoring speech sound disorders in children. The study introduces a novel forced alignment tool that leverages wav2vec 2.0 and UnsupSeg models. It utilizes class regions from wav2vec 2.0 labels to reference segments, employing maximal overlap to determine class labels. Post-processing steps further enhance segmentation performance, and self-supervised models reduce the need for labeled data. Evaluation on the TIMIT dataset demonstrates competitive phoneme segmentation results.

Tamiru and Abate [37] address the need for extracting both audio file structure and linguistic content from large archives, focusing on adding sentence boundaries to automatic speech transcription for Amharic. They present an automatic sentence-level speech segmentation system tailored for this language, using read and spontaneous speech corpora. Forced alignment, featuring monosyllable and tri-syllable acoustic models, along with rule-based and AdaBoost techniques, enhances accuracy. Evaluation results show promising performance, particularly with monosyllable acoustic models, making it valuable for processing Amharic speech archives.

Mahr et al. [34] investigates the performance of five forced-alignment algorithms in aligning phonemes in child speech. The study aims to provide insights for researchers working with child speech data, where manual segmentation is labor-intensive. The Montreal Forced Aligner with speaker adaptive training demonstrates the highest accuracy and the smallest timing differences, making it a reliable tool for researchers, especially when used in a semi-automated workflow with manual alignment inspection.

Leinonen et al. [38] explores challenges and solutions in creating speech corpora for under-resourced languages. It addresses biases in speech recognizers and methods to remove silences. Additionally, it breaks long interview audios into shorter segments and assesses the correlation between source and target language performance. The paper highlights tools like the Montreal Forced Aligner (MFA) and discusses evaluation methods for alignment accuracy, using Finnish and Komi.

Peng et al. [39] focuses on enhancing automatic phonetic segmentation in continuous speech. It identifies errors related to specific phonemes and boundaries and proposes strategies to improve segmentation accuracy, addressing intra-phone dissimilarity and inter-phone similarity challenges. This work offers practical methods to enhance automatic phonetic segmentation performance.

Punnoose [40] quantifies alignment skewness and investigates alignment errors in HMM-based forced alignment, revealing inherent skewness and significant differences in clean vs. noisy speech and other conditions. This study provides valuable insights into the biases of the Kaldi forced aligner.

Backstrom et al. [41] tackle forced alignment in sung speech, a distinct challenge from spoken speech. They employ a deep neural network to extract phonemic labels from sung acoustic signals, advancing forced alignment techniques in the context of sung speech analysis.

Milne [42] enhances forced alignment accuracy for variable pronunciations of word-final consonant clusters in French. It explores strategies such as model complexity adjustment, speaker-adaptive training, and dictionary refinement to improve alignment quality.

Rahmatullah and Ruan [43] evaluates HMM-based forced alignment for Indonesian using the Montreal Forced Aligner. It recommends a training strategy and highlights the importance of speaker diversity for alignment accuracy.

These researches collectively contribute to the evolving landscape of forced alignment research, addressing various challenges and proposing innovative solutions across different languages and domains.

C. REVIEW OF EXISTING ARABIC SEGMENTATION STUDIES

The inherent complexities and distinct features of the Arabic language, along with the intricate Tajweed rules employed in Qur'anic recitation, have posed unique challenges for researchers. These intricacies have not only driven but also necessitated the creation of specialized systems explicitly tailored for the segmentation of Arabic speech. This subsection will delve into these research efforts, illuminating their significance and contributions to the broader field of speech segmentation within the domain of Arabic language processing.

Al-Manie et al. [44] employed the zero crossing rates (ZCR) algorithm to segment Arabic speech into various categories such as voiced/unvoiced fricatives, stops, nasals, laterals, and trills. Their study, conducted on the KACST dataset, yielded favorable results compared to manual segmentation.

Awais et al. [10] presented an Arabic phoneme segmentation algorithm utilizing fast Fourier transform (FFT) spectrogram analysis. They employed intensity and phoneme duration to distinguish pauses from consonants, achieving a segmentation accuracy of 95.39% on a continuous Arabic speech dataset with 10 male speakers and a total of 2346 phonemes.

Anwar et al. [45] introduced a Qur'anic phoneme segmentation system incorporating cues such as zero crossing rate (ZCR), power spectral density (PSD), formant transitions, consonant and vowel rhythm, intonation patterns (fundamental frequency), and vowel duration. Their method demonstrated an accuracy of up to 89% on continuous speech files from eight different speakers.

Nofal et al. [46] presented a forced alignment method employing the Viterbi algorithm for automatic segmentation using a Non-Segmented data trained Acoustic model (NSAM) in Arabic. The NSAM, created by training various Arabic acoustic models with HMM-based HTK ver 3.1 training tools, was used as input for an acoustic modeling-based segmentation system, generating phonetically segmented data. This led to the development of an Automatically Segmented data trained Acoustic Model (ASAM).

Tolba et al. [23] proposed an algorithm based on wavelet transform and spectral analysis for consonant and vowel (C/V) segmentation in Arabic speech without linguistic information. Their focus was on identifying transitions between consonant and vowel segments using wavelet packet decomposition, achieving an accuracy rate of approximately 88.3% for C/V segmentation in a set of 20 Arabic words.

Peng et al. [39] conducted experiments within the classic GMM-HMM framework to address issues related to the diphthong /oy/ and glide-to-vowel boundaries in Arabic speech. They introduced two coupled approaches involving GMM and context-dependent models to enhance automatic phonemic segmentation.

Absa et al. [47] developed a hybrid speech segmentation algorithm using Genetic Algorithm (GA) optimization over multiple features, including entropy, zero crossing, and energy. Their results showed good accuracy compared to manual segmentation using the KACST database.

Karim [48] proposed three independent Genetic Algorithms (GAs) for optimizing the parameters of Indonesian syllable speech segmentation, focusing on boundary detection, iterative splitting, and iterative assimilation.

Frihia and Bahi [49] suggested a combination of HMMs and Support Vector Machines (SVMs) for waveform phoneme segmentation and labeling, which notably improved speech recognizer accuracy by approximately 0.05%, even in the presence of background noise.

Ahcène et al. [50] introduced an automatic segmentation system for Arabic speech into phonemes using a combination of HMM and ANN, classifying speech signals into five classes: fricatives, plosives, nasals, liquids, and vowels.

Javed et al. [51] presents an innovative unsupervised phonetic segmentation method for Classical Arabic speech, emphasizing the detection of phoneme boundaries in speech utterances. The method leverages cosine distance similarity scores and combines forward and inverse characteristics of the vocal tract (FICV) to achieve segmentation. The research evaluates this approach on a dataset of Classical Arabic. Comparative experiments against state-of-the-art techniques, including hidden Markov model-based forced alignment, demonstrate that the FICV-based method excels with a 14.48% total error rate and an 85.2% alignment accuracy within a 10 ms alignment error. However, it's worth noting that there has been limited attention to the rules of the Medd, a challenging aspect of phonetic segmentation, particularly when applied to Qur'anic recitation.

Alqadasi et al. [52] presented a Rule-Based Phoneme Duration Algorithm to enhance phoneme classification in Qur'anic recitation. They analyzed a Qur'anic dataset containing 21 Ayats recited by 30 individuals using a baseline Hidden Markov Model (HMM)-based speech recognition model. By optimizing phoneme classification models based on duration, their proposed algorithm achieved remarkable accuracy ranging from 99.87% to 100%, depending on the Medd type.

In conclusion, the body of research on Arabic speech segmentation can be categorized into two primary approaches: blind segmentation and aided segmentation algorithms. Blind segmentation relies on feature extraction methods to identify changes in speech signals independently of external linguistic knowledge. Aided segmentation, on the other hand, incorporates linguistic data like orthographic or phonetic transcription into the algorithm's training process, often employing classification techniques to enhance segmentation accuracy.

While significant progress has been achieved in the advancement of Arabic speech segmentation methods, the distinctive characteristics of Qur'anic recitation, especially the elongation of long vowels, persist as formidable challenges. This ongoing challenge both inspires researchers and provides an opportunity for further development and enhancement of speech segmentation techniques tailored to Qur'anic recitation.

III. THE PROPOSED AUTOMATIC SPEECH SEGMENTATION

This research aims to develop an automated speech segmentation approach tailored explicitly for Qur'an recitation. This involves refining the HMM-based forced alignment algorithm by incorporating a specialized acoustic model designed for the Qur'an. Given the distinct characteristics of Qur'anic recitation, our focus lies in distinguishing it from standard Arabic speech, particularly concerning the Tajweed rules that intricately govern phoneme duration.

The primary challenge in developing an automatic speech segmentation system for Qur'an recitation stems from the intricate variations introduced by Tajweed rules. Tajweed serves as the guiding framework for the correct pronunciation and phonetic attributes of Qur'anic verses. It adds layers of complexity to the recitation, introducing unique nuances in phoneme duration, articulation, and pronunciation that are not encountered in everyday Arabic speech. Moreover, the absence of a comprehensive phonetic transcription of Qur'anic recitation within standard Arabic datasets poses another significant obstacle. Due to the Tajweed rules that govern Qur'anic recitation, there are variations in some phonetics and the manner of recitation compared to ordinary spoken Arabic. The Qur'anic recitation adheres to strict rules regarding the pronunciation and the duration of pronunciation, including the rules of Medd and Ghunnah. Consequently, researchers are compelled to assemble dedicated Qur'anic datasets and manually undertake the meticulous tasks of segmenting and labeling audio recordings. Existing datasets do not sufficiently encompass the full spectrum of phonetic variations present in Qur'anic recitation. Therefore, this research tackles the challenge of bridging the divide between the subtleties of Qur'anic phonetics and the limitations of standard Arabic datasets, offering a specialized solution tailored to this distinctive domain.

As a result, this project encompasses the compilation of Qur'anic datasets, along with the processes of manual segmentation and labeling, to address this issue. The proposed system encompasses several key stages, as illustrated in Figure 1: dataset preparation, the construction of an acoustic model utilizing the HMM toolkit (HTK), the extraction of speech segments via the Forced Alignment algorithm (FA), and an enhanced Forced Alignment procedure based on the dedicated Qur'anic acoustic model. Each of these steps will be meticulously examined and discussed in detail.

A. DATASET PREPARATION

The initial phase in constructing a phonetic segmentation system utilizing forced alignment algorithms involves sourcing a database that pairs audio recordings with precise phonetic transcriptions. This is because these algorithms heavily rely on knowing the precise timing of phonemes or words in the spoken audio. Consequently, the TIMIT database has become a popular choice among researchers. This inclination is rooted in the fact that, in contrast to TIMIT, the majority of extensive speech corpora lack detailed phonetic transcriptions [53]. When only transcriptions at the word level are available, forced alignment becomes notably more challenging. In such scenarios, the algorithm faces a greater task in accurately delineating phonemic boundaries since it lacks the fine-grained guidance provided by phonetic transcriptions. This underscores the importance of comprehensive phonetic annotations within speech databases, as they greatly facilitate precise forced alignment, consequently benefiting a wide array of applications in speech processing and analysis.

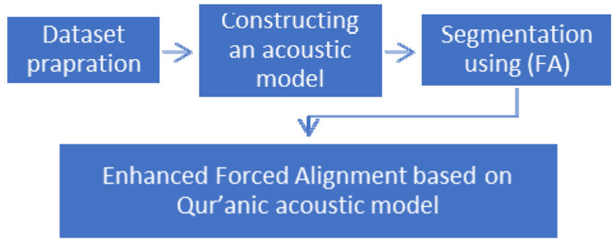

FIGURE 1. Flow chart for the proposed algorithm (Enhanced Forced Alignment based on Qur'anic acoustic model).

The absence of a dedicated speech corpus for Quranic recitation, especially one suitable for phoneme segmentation, presented a significant challenge [54], [55]. Existing Quranic databases primarily contain recitations from a limited number of speakers and often focus on specific surahs, thus limiting their usefulness for applications that require the recognition of all Tajweed rules [54]. To address this issue, we initiated the preparation of our database, which involved procuring speech data from Hafiz reciters (individuals who have memorized the Quran) and placing a specific emphasis on phoneme segmentation and labeling.

To prepare a speech corpus tailored for Quranic recitation, an existing database referenced as [56] was leveraged. This selection was a critical step in Quranic recitation recognition.

A pivotal element in constructing this corpus was the careful selection of 21 verses (Ayats) from diverse surahs to ensure the representation of fundamental Tajweed rules. These selected Ayats (shown in Table1) covered various Medd types, including Medd 'asli, Medd Wajib Mutasil, Medd Jayiz Munfasil, Medd Lazim Kalimi, Medd Lazim Harfi, and Medd Earid Lilsukun. This approach guaranteed diversity in Quranic recitation pronunciation and captured the full phoneme spectrum present in the Arabic language.

To obtain a comprehensive dataset for training and testing the Holy Quran recitation recognition model, two methods were employed: web-based data collection and live

TABLE 1. The verses selected to the dataset.

Surat:Ayah	Ayat	Medd Type
96:3	اقْرَأْ وَرَبُّكَ الْأَكْرَمُ	No Medd
108:2	فَصَلِّ لِرَبِّكَ وَانْحَرْ	No Medd
112:1	قُلْ هُوَ اللَّهُ أَحَدٌ	Medd 'asli
108:3	إِنَّ شَانِئَكَ هُوَ الْأَبْتَرُ	Medd 'asli
86:3	النَّجْمُ الثَّاقِبُ	Medd 'asli
86:4	إِنْ كُلُّ نَفْسٍ لَمَّا عَلَيْهَا حَافِظٌ	Medd 'asli
40:1	حم	Medd Lazim Harfi
50:1	ق وَالْقُرْآنِ الْمَجِيد	Medd Lazim Harfi
68:1	ن وَالْقَلَمِ وَمَا يَسْطُرُونَ	Medd Lazim Harfi
1:7	صِرَاطَ الَّذِينَ أَنْعَمْتَ عَلَيْهِمْ غَيْرِ الْمَغْضُو□ِ عَلَيْهِمْ وَلَا الضَّالِّينَ	Medd Lazim Kalimi
69:1	الْحَاقَّةُ	Medd Lazim Kalimi
39:64	قُلْ أَفَغَيْرَ اللَّهِ تَأْمُرُونِّي أَعْبُدُ أَيُّهَا الْجَاهِلُونَ	Medd Lazim Kalimi
86:11	وَالسَّمَاءِ ذَاتِ الرَّجْع	Medd Wajib Mutasil
89:23	وَجِيءَ يَوْمَئِذٍ بِجَهَنَّمَ	Medd Wajib Mutasil
27:11	إِلَّا مَن □َلَمَ ثُمَّ بَدَّلَ حُسْنًا بَعْدَ □ُوءٍ فَإِنِّي غَفُورٌ رَّحِيمٌ	Medd Wajib Mutasil
108:1	إِنَّا أَعْطَيْنَاكَ الْكَوْثَرَ	Medd Jayiz Munfasil
77:29	انطَلِقُوا إِلَى مَا كُنتُم بِهِ تُكَذِّبُونَ	Jayiz Munfasil + Sila Sughraa
106:4	الَّذِي أَطْعَمَهُم مِّن جُوعٍ وَآمَنَهُم مِّنْ خَوْفٍ	Medd Jayiz Munfasil + Medd Badal)
79:34	فَإِذَا جَاءَتِ الطَّامَّةُ الْكُبْرَى	'asli + Wajib +Lazim Kalimi+'asli
80:33	فَإِذَا جَاءَتِ الصَّاخَّةُ	Medd 'asli + Lazim Kalimi
78:36	جَزَاءً مِنْ رَبِّكَ عَطَاءً حِسَابًا	Wajib + Wajib + Medd Eiwad

recording. The web-based source provided a substantial volume of data quickly and efficiently by capturing the recitations of 100 renowned reciters. However, it lacked diversity in terms of recitation speed, the sex of reciters, and recording environments. In contrast, live recording offered the advantage of selecting reciters, recording settings, and diversity but required significantly more time and effort. When selecting reciters, their proficiency in Quranic recitation was a key criterion, ensuring that all selected reciters held professional certification in Quranic recitation.

The resulting dataset comprised 21 Ayats, totaling 19,800 words, with 89 unique words. The database encompassed seven hours and twelve minutes of audio. Each speaker contributed 567 phonemes in one session of recitation, totaling 40 unique phonemes out of 124,740 phonemes across all Ayats. The frequency distribution of phonemes is illustrated in Figure 2.

For audio segmentation at the Ayah level and preprocessing, the Audacity 2.3.2 audio processing tool was utilized, and the sampling rate of the segmented phoneme level in our study was 16 kHz, as shown in Figure 3. This dataset served as the foundation for training and testing the HMM-based acoustic model, with an 80% training and 20% testing split.

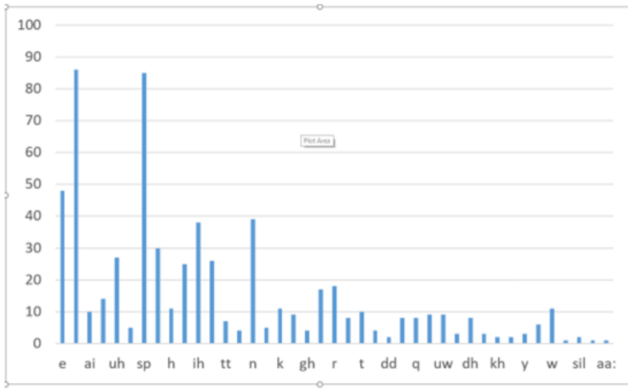

FIGURE 2. Frequency distribution for the phonemes.

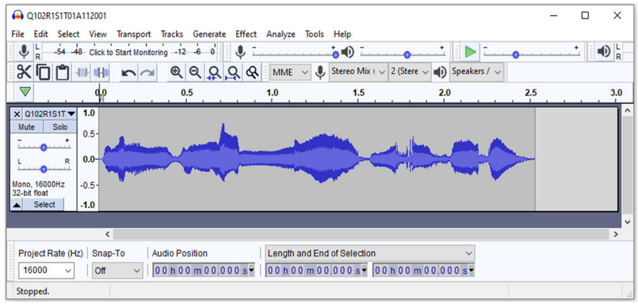

FIGURE 3. Employing audacity software for recording and preprocessing.

To evaluate the dataset, 27 words were selected. These words covered various proverbs from different verses at different recitation speeds and belonged to six random speakers, including three males and three females. Each reciter's recitations were recorded three times at different recitation speeds: Tawasut (Normal recitation rate n-speech), Hader (fast recitation rate f-speech), and Tahqiq (slow recitation rate s-speech). Consequently, 486 selected words (27 words $*$ 6 reciters $*$ 3 recitations) were manually segmented and labeled on the phoneme level for evaluation. When selecting phonemes, we focused on factors that could affect phoneme-level segmentation, including phoneme type (consonant, short vowel, and long vowel) and Tajweed rules related to phoneme duration, such as Medd, Ghunnah, and Tashdid. Praat tool software was utilized to perform the analysis, manual segmentation, and labeling at three levels: word, letter, and phoneme, as exemplified in Figure 4.

It is worth noting, as depicted in Figure 4, that some words contain silence-like parts, attributed to specific phonetic features known as 'plosives.' These silence-like parts are associated with some Arabic phonemes such as q, k, t, d, b. These silence-like parts are considered part of the phoneme due to their connection with the phonetic characteristics of certain Arabic letters.

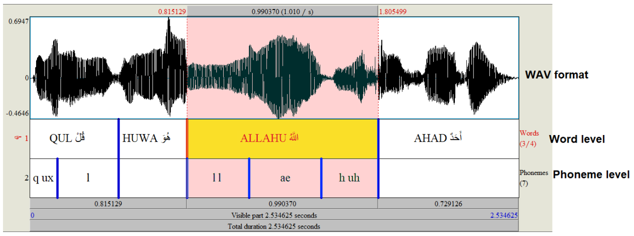

FIGURE 4. Example of Praat tools used for segmentation and labelling.

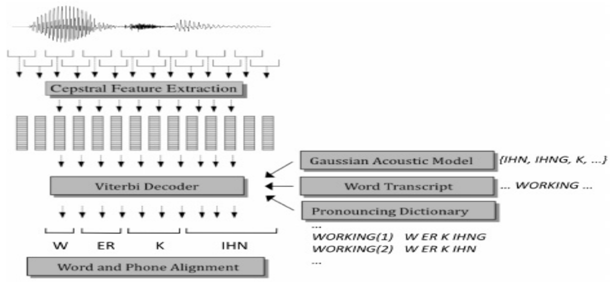

FIGURE 5. HTK results or Qur'anic Tajweed acoustic model.

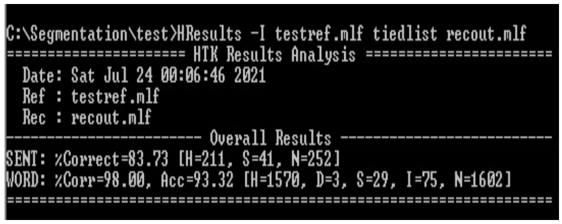

FIGURE 6. HMM-based Forced Alignment algorithms.

B. CONSTRUCTING AN ACOUSTIC MODEL USING HMM TOOLKIT (HTK)

This study involves building an acoustic model specifically designed for recognizing Quranic recitation using the HTK toolkit. The process comprises five key steps:

- Task Grammar: This involves creating a language model that consists of a comprehensive list of words and their probabilities of occurring in a given sequence. For Quranic recitation recognition, the word sequence is predetermined, following the Quran's specific order, but allowances must be made for pauses and repetitions.
- Dictionary: The next step is to construct a pronunciation dictionary. This involves generating a sorted list of words from the grammar, each with its associated pronunciation. Arabic phonetic dictionaries were used for this purpose, enabling HTK to compile an Acoustic Model.
- Transcription Files: The data needs to be transcribed at both word and phone levels since HTK cannot directly process the prompts file. Word-level transcriptions are accomplished using the Master Label File (MLF) approach, which contains transcriptions in a specific format. These word-level transcriptions are then expanded to phone-level transcriptions by replacing each word with its corresponding phoneme, generating a new Phone Level Master Label File.
- Feature Extraction: Audio files are converted from WAV format to Mel Frequency Cepstral Coefficient (MFCC) format to extract features that represent the speech.
- Triphone HMMs: The final stage involves creating context-dependent triphone Hidden Markov Models (HMMs). This includes two steps: generating triphones from monophones and creating tied-state triphones.

For evaluation purposes, the Word Error Rate (WER) was measured using the HTK tool called "HResults." The recognition model's performance was assessed against the training data collected. The results indicate that 83.73% of the test sentences were correctly recognized, with 98% of words recognized accurately. However, due to insertion errors (recognized words not present in the audio file), the accuracy rating for words recognized drops to 93.32%. The "D" represents Deletion Errors, "S" signifies Substitution Errors, and "I" indicates Insertion Errors. Figure 5 shows the result of the recognition accuracy assessment for the line starting with SENT.

C. SPEECH SEGMENTS USING HMM FORCED ALIGNMENT ALGORITHM (FA)

Forced alignment is the automated process of aligning orthographic transcriptions with specific time intervals in audio files to create segmentations at the sentence, word, or phone level.

Figure 6 illustrates the concept of forced alignment. This procedure pertains to the category of algorithms that simultaneously ingest an audio file and its associated transcription as input elements. It calculates precise time points in the audio file that correspond to each word and even each phoneme indicated in the transcription [57]. These algorithms undergo training using Hidden Markov Models (HMMs), which learn by applying temporal frames to the audio recording, extracting spectral information like formants and wave intensity for each phoneme. Following this training, these frames are used to traverse the signal and identify potential transition points between two phonemes [58]. The forced alignment algorithm relies on MFCC-Based Cosine Distance Analysis. To assess the viability of MFCC as a feature for speech segmentation, it applies thirty triangular filter banks on a mel scale to transform the power spectrum of the speech. The resulting energies from each filter bank are further processed to obtain their MFCC-based representation [51].

This research employs both categories of segmentation. Forced alignment, using HTK tools, segments the dataset

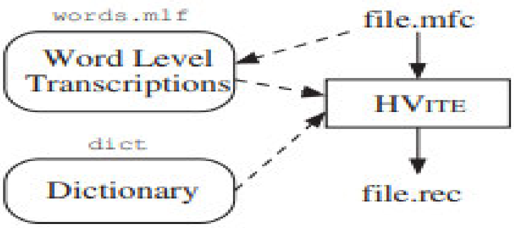

FIGURE 7. The process of Forced Alignment Utilizing HTK [59].

'/Q102R1S1T01A086004.rec

0	700000	sil	-535.8078	SENT-END
700000	1000000	e	-236.596863	IN
1000000	1700000	ih	-515.19397	
1700000	7900000	n	-3427.99707	
7900000	7900000	sp	-0.167745	
7900000	10000000	k	-1488.652222	KULLU
10000000	11100000	uh	-681.996643	
11100000	14000000	l	-1519.116943	
14000000	15100000	l	-751.104431	
15100000	16200000	uh	-674.936035	
16200000	16900000	sp	-522.538391	

(a)

	SENT-END
0	SENT-END
0.07	IN
0.79	KULLU
1.69	NAFSIN
2.3	LAMMA
3.73	AAALAYHA
4.87	HAFITH
5.95	SENT-END
0	SENT-END

(b)

FIGURE 8. Sample of the forced alignment output file (a) phonetic segment (b) word segment.

at both the word and phoneme levels. Additionally, manual segmentation is performed at both levels to facilitate the evaluation of the automated method.

HTK tools are utilized to create a new network for each input utterance based on word-level transcriptions and a lexicon. The resulting output transcription typically includes only the words and their boundaries. HTK is then employed in forced alignment mode to select the best-matching pronunciation. Subsequently, the HMMs can be retrained using new phone-level transcriptions. It's important to note that training data may contain periods of leading and trailing silence, necessitating the inclusion of a silence model at the start and end of the recognition network [59]. Figure 7 provides an overview of the core steps involved in the forced alignment design, with references to relevant files used in the process.

Figure 8 presents the output file generated by the forced alignment algorithm at both the phoneme and word levels. The score associated with each model name represents the acoustic score for that segment. In Figure 10, alongside each word, the score indicates the starting time of the word in milliseconds.

The efficacy of forced alignment is gauged through the computation of the Correct Classification Rate (CCR), denoted as:

This metric quantifies the percentage of frames that are correctly categorized. In this context, a frame is deemed accurately classified when it aligns with the same class as in manual segmentation [50].

In the HMM-based forced alignment approach, a common practice entails the representation of each sound using a three-state Hidden Markov Model (HMM) without permitting transitions to be skipped, meaning transitions can only occur to the next state or the eigenstate. The precise configuration of the HMM, coupled with the frame offset, dictates the minimum duration required to represent a sound. An earlier study has suggested a minimum duration of at least 30 milliseconds for precise phoneme identification [39].

Nonetheless, despite the efficacy of forced alignment in phoneme alignment for other languages, experiments have unveiled its limited accuracy in segmenting Quranic recitations. Specifically, it achieved an accuracy rate of only 7% within a 30-millisecond tolerance region when segmenting verses containing long vowels.

$$\text{HTK Forced alignment CCR\%} = \frac{23}{324} * 100 = 7\%$$

Upon scrutinizing the results, it became evident that phonemes such as Medd and Ghunnah exerted the most substantial influence on the suboptimal segmentation accuracy. For words in verses with long vowels, the segmentation accuracy (CCR) ranged from 0% to 28%, averaging 7% within a 30-millisecond tolerance region. In contrast, the CCR segmentation accuracy for words in verses devoid of long vowels could reach 94%, as exemplified by the word “قُلْ” in the conducted experiments.

To delve into the data and elucidate the causes behind the suboptimal segmentation accuracy of Quranic recitations, the dataset was categorized into three classes: Class 1 comprising verses without long vowels, Class 2 comprising verses with long vowels representing Medd ‘Asli, and Class 3 comprising verses with long vowels representing Medd Farei in accordance with Tajweed rules.

The word selection was comprehensive, encompassing factors that could influence phoneme segmentation in Quranic recitation, including the qualities of letters known in intonation, such as Hams (whispering), Al-Rakhawah (softness), Qalqalah, Ghunnah, and Tashdeed. An analysis of Figure 9, which displays the performance of the forced alignment method for phoneme segmentation in verses without Medd (long vowels), reveals high accuracy. Most segmentation ratios within the 30-millisecond tolerance, with an average of 57%, while the average CCR reached 66%, 84%, and 100% within tolerances of 50 ms, 70 ms, and 100 ms, respectively. Furthermore, it's apparent that letters with Haraka (short vowels) exhibit higher segmentation accuracy compared to Sakinah consonants, particularly the hamza and Qalqalah letters. Notably, Qalqalah letters displayed the lowest accuracy at 28%.

Figure 10 illustrates the performance of the forced alignment method when applied to verses containing “Asli Medd” (lengthening by 2 vowels). Most CCR values for phoneme segmentation are concentrated between 0.1 and 1 second, with CCR within 30 milliseconds ranging from 0% to 28%,

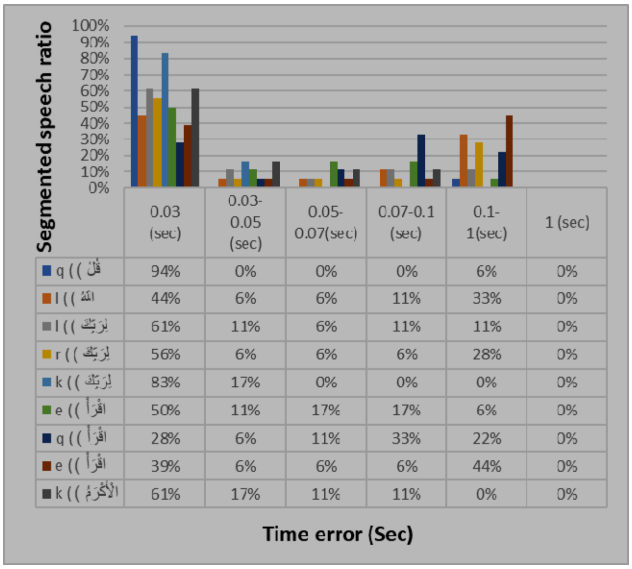

FIGURE 9. Performance of the phoneme segmentation using FA (Dataset class 1).

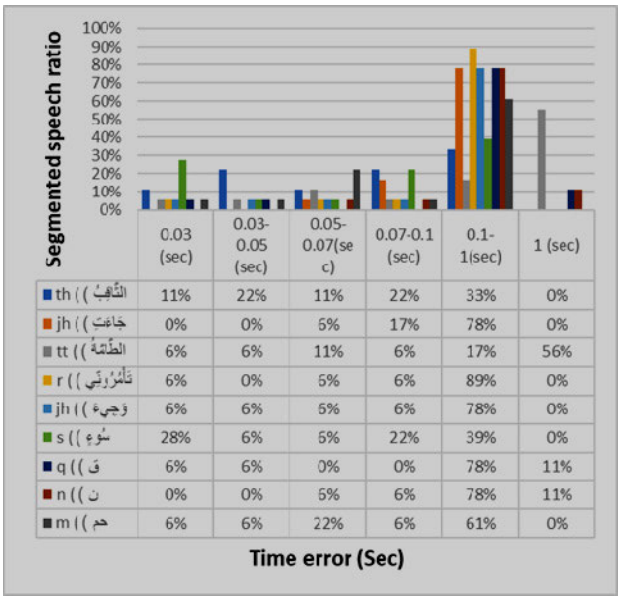

FIGURE 10. Performance of the phoneme segmentation using FA (Dataset class 2).

with the average CCR reaching 8%, 13%, 22%, 32%, and 93% within tolerances of 30 ms, 50 ms, 70 ms, and 100 ms, respectively.

Finally, Figure 11 shows the performance of the same method (forced alignment) when applied to verses containing one or more "Fare'i Medd" (lengthening by 4 or 6 vowels). The majority of CCR values for phoneme segmentation are observed beyond one second, with CCR within a 30-millisecond tolerance at 0% in most cases and reaching 22% in some instances, with the average CCR reaching 6%, 11%, 13%, 16%, and 69% within tolerances of 30 ms, 50 ms, 70 ms, and 100 ms, respectively.

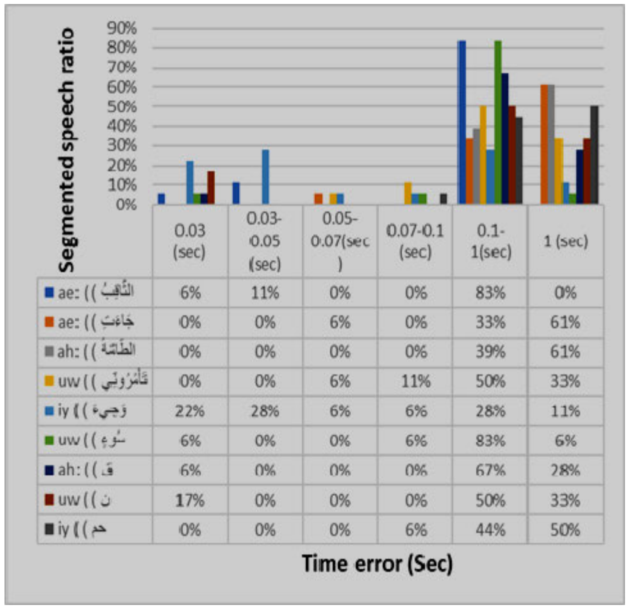

FIGURE 11. Performance of the phoneme segmentation using FA (Dataset class 3).

In summary, the characteristics of letters significantly impact phoneme segmentation in Quranic recitation, with the types of lengthening governed by Medd rules being particularly distinctive in Quranic recitation compared to ordinary speech.

D. QUR'AN ACOUSTIC MODEL-BASED FORCED ALIGNMENT SCHEME

To enhance the precision of phoneme alignment in the context of Qur'anic recitation, a proposed algorithm based on the acoustic model of Qur'anic recitation has been introduced. The recitation rate and Tajweed rules play a pivotal role in this approach. The proposed scheme is illustrated in Figure 12.

The primary aim of this scheme is to improve the forced alignment process for segmenting phonemes within Qur'anic recitation. It leverages the segmented word lengths generated by the forced alignment algorithm to compute the average duration of Harakah (short vowels) in the recitation. This average duration, in turn, serves as a critical factor in computing the recitation rate. The proposed algorithm further categorizes phonemes based on their Harakah duration and calculates the duration of all phonemes in accordance with the Qur'an acoustic model.

Phonemes characterized by durations of less than three Harakah (equivalent to three short vowels) are computed based on the established recitation rate. In contrast, phonemes with durations exceeding three Harakah are calculated by dividing the remaining length of the corresponding word, as obtained through the forced alignment, by the number of Harakah present. This approach facilitates the measurement of long vowels and Ghunnah (nasalization) following the computation of consonants with short vowels

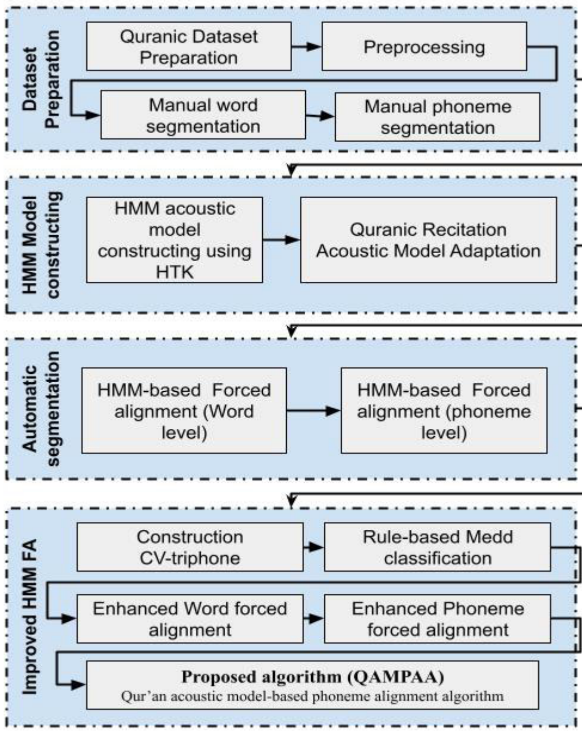

FIGURE 12. Basic flow for designing Qur'an Acoustic Model-Based Forced Alignment Scheme.

(Fathah, Kasrah, or Dhummah). The determination of word and phoneme weights relies on the phoneme classification algorithm developed and elucidated by Alqadasi, et al. [52] for Quranic recitation.

To execute this algorithm, a critical initial step involves word separation and length computation. This necessitates the processing of a wave file and its associated transcription using a forced alignment algorithm implemented through HTK, thereby obtaining segmented words. Subsequently, the first Enhanced Phoneme Forced Alignment (EPFA) is performed to enhance the accuracy of word segmentation. This proposed algorithm, which enhance phoneme segmentation, is referred to as EPFA.

The subsequent pseudocode outlines the critical component of the phoneme forced alignment optimization algorithm's processing.

Figure 13 illustrates the fundamental framework of the Enhanced Phoneme Forced Alignment (EPFA). Despite the initial enhancements introduced in the Enhanced Phoneme Forced Alignment (EPFA), it was observed that there still existed issues of accuracy in segmentation, particularly stemming from inadequate word-level segmentation within forced alignment for specific cases. Consequently, an algorithm was proposed to refine the alignment at the verse level, and it was named Enhanced Word Forced Alignment (EWFA).

As indicated in Figure 14, the proposed EWFA algorithm closely resembles the previous one, with a notable difference in step 17. This step involves the computation of phoneme

Algorithm

```
BEGIN
  Word_length = input_from_FA_output
  seg_Harakah = []
  X = []
  Harakah_length = 0
  w = Number_of_Words
  input_CV_triphone = read_file()
  X = compute_the_Harakah_Number_of_words()
  for i = 0 to w
   IF length(w) - i > 0 then
    X(i) = Max(Harakah_in_word)
    IF X(i) < X(i-1) then
     X + = [X(i)]
    ELSE
     Harakah_length = Word_length / X
     continue
  For j = 1 to w // Number_of_words
    For each phoneme in input_CV_triphone(w(J)) // word by word
     Length1 = []
     Length2 = []
     if (phoneme in input_CV_triphone Harakah < 3)
      seg_Harakah + = input_CV_triphone Harakah * Harakah_length
      Length1 + = seg_Harakah
    ELSE
      seg_Harakah + = (Word_length - Length1) / sum (Harakah > 3)
      * CV_triphone Harakah
     continue
    ELSE:
      seg_Harakah + = seg_Harakah
     END
    END
   END
  END
```

durations with more than three Harakah at the verse level. Essentially, it calculates the cumulative time of phonemes with fewer than three Harakah and then distributes the remaining time within the verse among the phonemes with more than three Harakah.

The subsequent pseudocode delineates the pivotal aspect of the word forced alignment optimization algorithm's processing.

The last enhancement involves a hybrid approach that combines elements from EPFA and EWFA, leveraging the Qur'anic acoustic model to enhance the accuracy of HTK forced alignment as shown in Figure 15. This algorithm is referred to as the Qur'an Acoustic Model-Based Phoneme Alignment Algorithm (QAMPAA).

IV. EXPERIMENTAL RESULTS AND DISCUSSION

The segmentation algorithm's performance is assessed using a dataset of 324 test samples. To evaluate the accuracy of the segmented phonemes, a time error analysis is conducted, defined as:

$$\begin{aligned}\text{Time Error} \\ &= |\mathit{duration\ of\ manual\ segment\ boundary} \\ &\quad - \mathit{duration\ of\ the\ automatic\ segment\ boundary}\end{aligned}$$

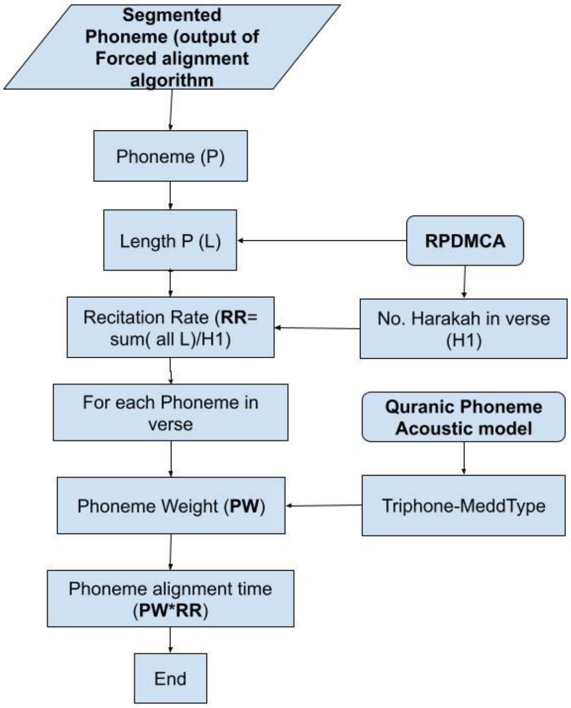

FIGURE 13. Displays the essential steps of the EPFA design.

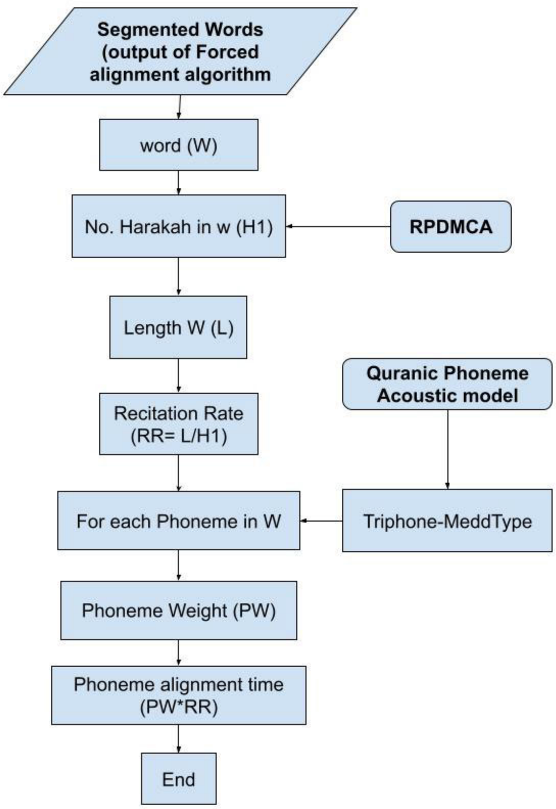

FIGURE 14. Displays the essential steps of the EPFA design.

Table 2 presents sample of segmented output results and the corresponding error rates for both the proposed algorithm and the classic HMM forced alignment. This

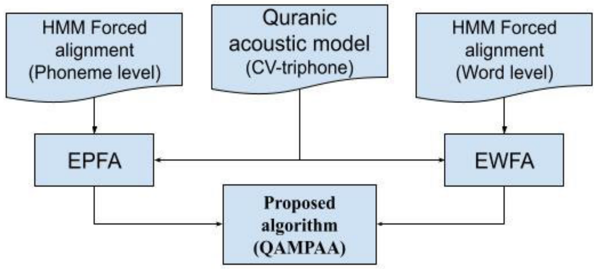

FIGURE 15. Hybrid of the EFA1 and EFA2 to obtain the Qur'an Acoustic Model-Based Phoneme Alignment Algorithm.

Algorithm

```
For j = 1 to triphone_p // Number of phonemes in CV-triphone
  For each phoneme in input_CV_triphone(J) // Phoneme by phoneme
    Length1 = []
    Length2 = []
    if (phoneme in input_CV_triphone Harakah < 3)
      seg_Harakah + = input_CV_triphone Harakah * Harakah_length
      Length1 + = seg_Harakah
    ELSE
      Length2 + = seg_Harakah
      seg_Harakah + = Length2 / sum(Harakah > 3) * CV_triphone Harakah
      continue
    ELSE:
      seg_Harakah + = seg_Harakah
    END
  END
END
```

TABLE 2. Sample of duration of segmented output by using manual segmentation, HTK forced alignment (FA), and Qur'an phonetic-based forced alignment algorithm (QAMPAA).

Word	Manual segmentation (sec)	FA (sec)	QAMPAA (sec)	Time error (FA) (sec)	Time error (QPAA) (sec)
الثَّاقِبُ	0.485	0.27	0.464	0.215	0.021
جَاءَتِ	1.429	0.03	1.418	1.399	0.011
الطَّامَّةُ	2.117	0.23	2.340	1.887	0.223
تَأْمُرُونِّي	1.489	1.551	3.037	0.062	1.548
وَجِيءَ	1.754	1.581	2.124	0.173	0.370
□ُوءٍ	1.490	1.411	2.102	0.079	0.612
ق	2.079	1.671	1.615	0.408	0.464
ن	1.335	0.37	1.057	0.965	0.278
حم	1.590	1.491	1.617	0.099	0.027

demonstrates the effectiveness of the proposed algorithm QAMPAA.

The experiments conducted highlight the superior performance of the QAMPAA, which aligns closely with manual segmentation, particularly in cases involving long vowels. In contrast, the forced alignment algorithm exhibits lower accuracy when dealing with long vowels, prompting us to emphasize long vowels in our development and evaluation process. Figure 16 illustrates the results of both classical

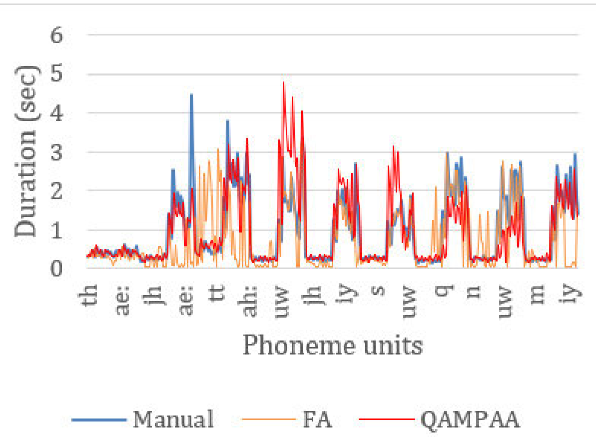

FIGURE 16. Sample of duration of phoneme units obtained by manual analysis and segmentation algorithm.

TABLE 3. Analysis of the Correct Classification Rate (CCR) across all automated segmentation methods.

Dataset class	Tolerance duration (sec)	0.03	0.05	0.07	0.1	1
class 1	FA	93	107	118	135	162
		57%	66%	73%	83%	100%
	EPFA	93	95	96	113	162
		57%	59%	59%	70%	100%
	EWFA	101	103	105	114	162
		62%	64%	65%	70%	100%
	QAMPAA	113	133	143	155	162
		70%	82%	88%	96%	100%
class 2	FA	11	20	33	49	148
		7%	12%	20%	30%	91%
	EPFA	52	83	93	103	161
		32%	51%	57%	64%	99%
	EPFA	51	81	97	115	162
		31%	50%	60%	71%	100%
	QAMPAA	81	113	124	137	162
		50%	70%	77%	85%	100%
class 3	FA	10	17	20	25	111
		6%	10%	12%	15%	69%
	EPFA	10	15	23	34	142
		6%	9%	14%	21%	88%
	EPFA	9	15	22	38	145
		6%	9%	14%	23%	90%
	QAMPAA	26	41	50	68	145
		16%	25%	31%	42%	90%

forced alignment algorithms and the proposed algorithm, comparing them to manual segmentation of all Arabic phonemes. Table 3 presents the performance results, offering a comparative view of the Correct Classification Rate (CCR) for the classical forced alignment, EPFA, EWFA, and the proposed QAMPAA algorithm. The test database was categorized into three groups: the first comprised seven words from verses without prolonged Medd (4 or 6 vowels), the second encompassed seven words from verses containing prolonged Medd, and the third involved seven words from verses with multiple prolonged Medd.

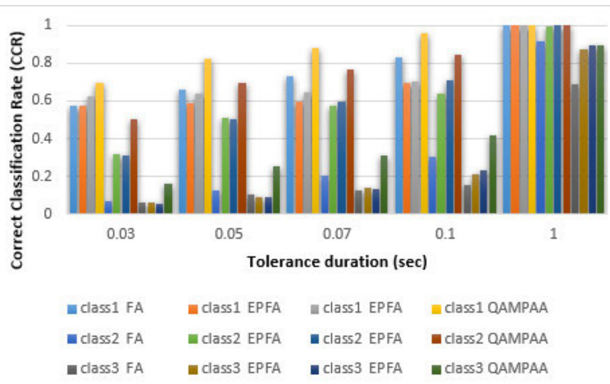

FIGURE 17. Comparison of segmentation CCR across different methods.

The results in Table 3 reveal that the classical forced alignment performs acceptably in the first category, achieving a CCR of 57% for verses without prolonged Medd. However, the CCR decreases significantly for the other two categories, both containing prolonged Medd, reaching as low as 6%.

Furthermore, Table 3 underscores the substantial improvement introduced by the proposed QAMPAA algorithm across all three categories. In the first category, the CCR stands at 70%, and in the second category, it reaches 50%. For the third category, the CCR rate shows improvement at 16%, although further enhancements are still needed in this category. Figure 17 refer to a visual representation of these improvements, which presents a comparison of all categories and the applied algorithms in this study.

The obtained result signifies the need for further experimentation and development of segmentation algorithms that specifically consider the unique characteristics of letters in Quranic recitation.

V. CONCLUSION

In this study, we present a series of enhancements aimed at refining the process of automatic phoneme segmentation, particularly in the context of Holy Quran recitation. To facilitate our research, we carefully curated a comprehensive dataset comprising 21 Quranic verses recited by 100 reciters. Additionally, 30 reciters were asked to record the same verses at varying recitation speeds. Leveraging the HTK toolkit, we established a phonetic model and employed a forced alignment algorithm for segmenting this dataset at both phoneme and word levels. Furthermore, we conducted meticulous pre-processing to adapt the model for Quranic recitation. The test sample underwent manual division at both word and phoneme levels to facilitate comparative analysis.

Our research endeavors have revealed a significant challenge in the realm of phoneme segmentation in Quranic recitation recognition. Traditional forced alignment methods, such as the Hidden Markov Model (HMM)-based algorithm, exhibit commendable accuracy when applied to Quranic verses devoid of extended vowels. However, their performance experiences a substantial decline when confronted with verses featuring these elongated vowel sounds.

In response to this challenge, we introduce a groundbreaking algorithm known as the Quranic Acoustic Model-Based Phoneme Alignment Algorithm (QAMPAA). This algorithm is meticulously crafted to cater to the unique and intricate phonetic characteristics of Quranic recitation. It comprehensively addresses aspects like intonation rules, the phenomenon of lengthening long vowels, and adherence to Tajweed guidelines. The implementation of the QAMPAA algorithm marks a pivotal turning point in the domain of audio segmentation, resulting in a remarkable boost in Correct Classification Rates (CCR) by 45% across diverse categories of Quranic verses. This substantial enhancement in segmentation accuracy is particularly evident within a 30-millisecond tolerance window.

Nonetheless, it's essential to acknowledge that our algorithm's performance can benefit from further refinement, particularly when dealing with Quranic verses characterized by complex phonetic structures and multiple extended vowels. While the QAMPAA algorithm achieves its highest CCR of 70% for verses without long vowels, indicating success in alignment tasks that don't involve these specific phonetic elements, there's room for improvement in addressing verses with more intricate phonetic patterns.

This paper offers valuable insights into the complexities of Quranic recitation recognition, making a significant contribution to various speech processing applications related to the Arabic language and Quranic recitation. These applications include, but are not limited to, speech recognition and speech synthesis. It underscores the importance of developing algorithms that are intrinsically attuned to the unique phonetic and intonational attributes of Quranic verses. The introduction of the QAMPAA algorithm marks a significant milestone in this pursuit, showcasing notable advancements in audio segmentation precision. However, it also underscores the need for ongoing refinement, especially concerning verses characterized by intricate phonetic structures. Future research endeavors should delve deeper into the distinct phonetic characteristics of Quranic recitation, with the goal of unlocking further enhancements in recognition and alignment accuracy.

ACKNOWLEDGMENT

The authors would like to thank the International Islamic University Malaysia (IIUM) and Universiti Teknologi Malaysia (UTM) for their invaluable assistance and for granting the opportunity to conduct this research. Additionally, the ITQAN Education Foundation (YAYASAN Pendidikan ITQAN) deserves special recognition for their indispensable support and resources, which played an instrumental role in the successful culmination of this study.

REFERENCES

[1] N. Jiang and T. Liu, "An improved speech segmentation and clustering algorithm based on SOM and K-means," *Math. Problems Eng.*, vol. 2020, pp. 1–19, Sep. 2020, doi: 10.1155/2020/3608286.

[2] M.-S. Salam and D. Mohamad, "Insertion reduction in speech segmentation using neural network," in *Proc. Int. Symp. Inf. Technol.*, Aug. 2008, pp. 1–7.

[3] M. Kalamani, S. Valarmathy, S. Anitha, and R. Mohan, "Review of speech segmentation algorithms for speech recognition," *Int. J. Adv. Res. Electron. Commun. Eng.*, vol. 3, no. 11, pp. 1–3, 2014.

[4] K. Geetha and E. Chandra, "Automatic phoneme segmentation of Tamil utterances," in *Proc. Int. Conf. Adv. Comput. Commun. Syst.*, Jan. 2015, pp. 1–4.

[5] S. He and H. Zhao, "Automatic syllable segmentation algorithm of Chinese speech based on MF-DFA," *Speech Commun.*, vol. 92, pp. 42–51, Sep. 2017, doi: 10.1016/j.specom.2017.04.003.

[6] H. Kamper, A. Jansen, and S. Goldwater, "Unsupervised word segmentation and lexicon discovery using acoustic word embeddings," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 24, no. 4, pp. 669–679, Apr. 2016, doi: 10.1109/TASLP.2016.2517567.

[7] L. Gan and Y. Zhang, "Investigating self-attention network for Chinese word segmentation," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 28, pp. 2933–2941, 2020, doi: 10.1109/TASLP.2020.3030487.

[8] W. Zhang, M. Pang, R. Du, and Y. Liu, "Automatic speech sentence segmentation from multi-paragraph databases," in *Proc. Int. Conf. Measuring Technol. Mechatronics Autom.*, vol. 1, Mar. 2010, pp. 721–724.

[9] D. Rybach, C. Gollan, R. Schluter, and H. Ney, "Audio segmentation for speech recognition using segment features," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Apr. 2009, pp. p4197–4200.

[10] M. Awais, S. Masud, and S. Shamail, "Continuous Arabic speech segmentation using FFT spectrogram," in *Proc. Innov. Inf. Technol.*, Nov. 2006, pp. p1–6.

[11] A. G. Adami and H. Hermansky, "Segmentation of speech for speaker and language recognition," in *Proc. 8th Eur. Conf. Speech Commun. Technol. (Eurospeech)*, Sep. 2003, pp. 1–4.

[12] T. T. Anh, M. N. Huu, and K. N. Trong, "A method for automatic vietnamese speech segmentation," *Int. J. Innov. Technol. Exploring Eng.* (IJITEE), vol. 8, no. 11, 2019, doi: 10.35940/ijitee.K2427.0981119.

[13] E. Saiegh-Haddad and R. Henkin-Roitfarb, "The structure of Arabic language and orthography," in *Handbook Arabic literacy*. Cham, Switzerland: Springer, 2014, pp. 3–28.

[14] A. M. Alashqar, "A comparative study on Arabic POS tagging using Quran corpus," in *Proc. 8th Int. Conf. Informat. Syst. (INFOS)*, May 2012, pp. NLP-29–NLP-33.

[15] A. Rushdi, A. R. Swayd, and K. Karul, *Tajweed Rules in the Holy Quran*. Syria-Damascus, Syria: Dar Al-Khair Islamic Books Publisher, 2000.

[16] K. C. Czerepinski and A. R. Swayd, *Tajweed Rules of the Qur'an*. Syria-Damascus, Syria: Dar Al-Khair Islamic Books Publisher, 2006.

[17] S.-L. Liew, "A large, curated, open-source stroke neuroimaging dataset to improve lesion segmentation algorithms," *Sci. Data*, vol. 9, no. 1, pp. 1–12, Jun. 2022, doi: 10.1038/s41597-022-01401-7.

[18] O. Scharenborg, V. Wan, and M. Ernestus, "Unsupervised speech segmentation: An analysis of the hypothesized phone boundaries," *J. Acoust. Soc. Amer.*, vol. 127, no. 2, pp. 1084–1095, Feb. 2010, doi: 10.1121/1.3277194.

[19] M. M. Rahman, F. Khatun, and M. A.-A. Bhuiyan, "Blocking black area method for speech segmentation," *Int. J. Adv. Res. Artif. Intell.*, vol. 4, no. 2, pp. 1–6, 2015.

[20] E. A. Kaur and E. T. Singh, "Segmentation of continuous Punjabi speech signal into syllables," in *Proc. World Congr. Eng. Comput. Sci.*, vol. 1, 2010, pp. 20–22.

[21] A. Hossain, N. Nahid, N. N. Khan, D. C. Gomes, and S. M. Mugab, "Automatic silence/unvoiced/voiced classification of Bangla velar phonemes: New approach," in *Proc. 8th ICCIT*, 2005.

[22] N. Sharma and P. Singh, "Automatic segmentation of wave file," *Int. J. Comput. Sci. Commun.*, vol. 1, no. 2, pp. 267–270, 2010.

[23] M. Tolba, T. Nazmy, A. Abdelhamid, and M. Gadallah, "A novel method for Arabic consonant/vowel segmentation using wavelet transform," *Int. J. Intell. Cooperat. Inf. Syst.*, vol. 5, no. 1, pp. 353–364, 2005.

[24] I. Mporas, T. Ganchev, and N. Fakotakis, "A hybrid architecture for automatic segmentation of speech waveforms," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Mar. 2008, pp. 4457–4460.

[25] B. Ziółko, S. Manandhar, R. C. Wilson, and M. Ziółko, "Wavelet method of speech segmentation," in *Proc. 14th Eur. Signal Process. Conf.*, Sep. 2006, pp. 1–5.

[26] S. M. Siniscalchi, P. Schwarz, and C.-H. Lee, "High-accuracy phone recognition by combining high-performance lattice generation and knowledge based rescoring," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2007.

[27] A. M. A. Al-Qadasi, M. S. Sunar, M. S. H. Salam, and R. Abdulghafor, "Medd recognition for Al-Quran recitation using ANN-based model," in *Proc. Post Pandemic Challenges Embracing Soc.*, 2023, p. 1.

[28] J. Kamarauskas, "Automatic segmetation of phonemes using artificial neural networks," *Elektronika IR Elektrotechnika*, vol. 72, no. 8, pp. 39–42, 2006.

[29] P. Bansal, A. Pradhan, A. Goyal, A. Sharma, and M. Arora, "Speech synthesis-automatic segmentation," *Int. J. Comput. Appl.*, vol. 98, no. 4, pp. 29–31, 2014.

[30] J. Dines, S. Sridharan, and M. Moody, "Automatic speech segmentation with HMM," in *Proc. 9th Austral. Conf. Speech Sci. Technol.*, 2002, pp. 544–549.

[31] A. Stolcke, N. Ryant, V. Mitra, J. Yuan, W. Wang, and M. Liberman, "Highly accurate phonetic segmentation using boundary correction models and system fusion," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2014, pp. 5552–5556.

[32] A. Baby, J. J. Prakash, and H. A. Murthy, "A hybrid approach to neural networks based speech segmentation," in *Proc. Int. Symp. Frontiers Res. Speech Music*, vols. 15–16. Rourkela, India: National Institute of Technology (NIT), 2017.

[33] J. Yuan, W. Lai, C. Cieri, and M. Liberman, "Using forced alignment for phonetics research," in *Chinese Language Resources and Processing: Text, Speech and Language Technology*. Cham, Switzerland: Springer, 2018.

[34] T. J. Mahr, V. Berisha, K. Kawabata, J. Liss, and K. C. Hustad, "Performance of forced-alignment algorithms on Children's speech," *J. Speech, Lang., Hearing Res.*, vol. 64, no. 6S, pp. 2213–2222, Jun. 2021, doi: 10.1044/2020_JSLHR-20-00268.

[35] H. Kim and H.-S. Choi, "Towards trustworthy phoneme boundary detection with autoregressive model and improved evaluation metric," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2023, pp. 1–5, doi: 10.1109/ICASSP49357.2023.10096748.

[36] B. Wohlan, D.-S. Pham, K. Y. Chan, and R. Ward, "A text-independent forced alignment method for automatic phoneme segmentation," in *AI 2022: Advances in Artificial Intelligence*, Cham, H. Aziz, D. Corrêa, and T. French, Eds. Cham, Switzerland: Springer, 2022, pp. 585–598, doi: 10.1007/978-3-031-22695-3_41.

[37] R. M. Tamiru and S. T. Abate, "Sentence-level automatic speech segmentation for amharic," in *Proc. 6th Int. Congr. Inf. Commun. Technol. (ICICT)*, vol. 2, London, U.K. Cham, Switzerland: Springer, 2022, pp. 477–485, doi: 10.1007/978-981-16-2380-6_42.

[38] J. Leinonen, N. Partanen, S. Virpioja, and M. Kurimo, "Semiautomatic speech alignment for under-resourced languages," in *Proc. Workshop Resour. Technol. Indigenous, Endangered Lesser-Resourced Lang. Eurasia Within 13th Lang. Resour. Eval. Conf.*, 2022, pp. 17–21. [Online]. Available: https://aclanthology.org/2022.eurali-1.3

[39] W. Peng, Y. Gao, B. Lin, and J. Zhang, "A practical way to improve automatic phonetic segmentation performance," in *Proc. 12th Int. Symp. Chin. Spoken Lang. Process. (ISCSLP)*, Jan. 2021, pp. 1–5.

[40] A. K. Punnoose, "A study on forced alignment error patterns in Kaldi," in *Proc. 8th Int. Conf. Signal Process. Commun. (ICSC)*, Dec. 2022, pp. 250–253, doi: 10.1109/ICSC56524.2022.10009308.

[41] D. A. Backstrom, B. V. Tucker, and M. C. Kelley, "Forced-alignment of the sung acoustic signal using deep neural nets," *Can. Acoust.*, vol. 47, no. 3, pp. 98–99, 2019.

[42] P. M. Milne, "Improving the accuracy of forced alignment through model selection and dictionary restriction," M.S. thesis, Dept. Linguistics, McGill Univ., 2015.

[43] G. M. Rahmatullah and S.-J. Ruan, "Performance evaluation of Indonesian language forced alignment using Montreal forced aligner," in *Proc. 6th Int. Symp. Comput., Consum. Control (IS3C)*, Jun. 2023, pp. 163–166, doi: 10.1109/IS3C57901.2023.00051.

[44] M. A. Al-Manie, "Arabic speech segmentation: Automatic verses manual method and zero crossing measurements," *Indian J. Sci. Technol.*, vol. 3, no. 12, pp. 1134–1138, Dec. 2010.

[45] M. J. Anwar, M. Awais, S. Masud, and S. Shamail, "Automatic Arabic speech segmentation system," *Int. J. Inf. Technol.*, vol. 12, no. 6, pp. 102–111, 2006.

[46] M. Nofal, E. Abdel-Raheem, H. El Henawy, and N. S. A. Kader, "Arabic automatic segmentation system and its application for Arabic speech recognition system," in *Proc. 46th Midwest Symp. Circuits Syst.*, Dec. 2003, pp. 697–700.

[47] A. H. A. Absa, M. Deriche, M. Elshafei-Ahmed, Y. M. Elhadj, and B.-H. Juang, "A hybrid unsupervised segmentation algorithm for Arabic speech using feature fusion and a genetic algorithm (July 2018)," *IEEE Access*, vol. 6, pp. 43157–43169, 2018, doi: 10.1109/Access.2018.2859631.

[48] R. M. Karim and Suyanto, "Optimizing parameters of automatic speech segmentation into syllable units," *Int. J. Intell. Syst. Appl.*, vol. 11, no. 5, pp. 9–17, May 2019.

[49] H. Frihia and H. Bahi, "HMM/SVM segmentation and labelling of Arabic speech for speech recognition applications," *Int. J. Speech Technol.*, vol. 20, no. 3, pp. 563–573, Sep. 2017.

[50] A. Ahcène, A. Aissa, D. Abdelkader, B. Khadidja, and D. Ghania, "Automatic segmentation of Arabic speech signals by HMM and ANN," in *Proc. Int. Conf. Electr. Sci. Technol. Maghreb (CISTEM)*, Oct. 2016, pp. 1–4.

[51] M. Javed, M. M. A. Baig, and S. A. Qazi, "Unsupervised phonetic segmentation of classical Arabic speech using forward and inverse characteristics of the vocal tract," *Arabian J. Sci. Eng.*, vol. 45, no. 3, pp. 1581–1597, Mar. 2020, doi: 10.1007/s13369-019-04065-5.

[52] A. M. A. Alqadasi, M. S. Sunar, S. Turaev, R. Abdulghafor, M. S. Hj Salam, A. A. S. Alashbi, A. A. Salem, and M. A. H. Ali, "Rule-based embedded HMMs phoneme classification to improve Qur'anic recitation recognition," *Electronics*, vol. 12, no. 1, p. 176, Dec. 2022.

[53] J. Yuan, N. Ryant, M. Liberman, A. Stolcke, V. Mitra, and W. Wang, "Automatic phonetic segmentation using boundary models," in *Proc. Interspeech*, Aug. 2013, pp. 2306–2310. [Online]. Available: https://www.microsoft.com/en-us/research/publication/automatic-phonetic-segmentation-using-boundary-models/

[54] A. Mohammed and M. S. Sunar, "Toward A rich quranic Arabic speech corpus for Tajweed rules," in *Proc. Universal Wellbeing (ICUW)*, 2019, p. 210.

[55] A. M. A. Alqadasi, R. Abdulghafor, M. S. Sunar, and M. S. B. H. J. Salam, "Modern standard Arabic speech corpora: A systematic review," *IEEE Access*, vol. 11, pp. 55771–55796, 2023, doi: 10.1109/ACCESS.2023.3282259.

[56] A. M. A. Al-Qadasi, "Phoneme duration scheme for tajweed medd rules recognition in Qur'an recitation," Ph.D. dissertation, Dept. Comput. Sci., Universiti Teknologi Malaysia, Malasia, 2021. [Online]. Available: http://eprints.utm.my/id/eprint/101526/

[57] C. W. Wightman and D. T. Talkin, "The aligner: Text-to-speech alignment using Markov models," in *Prog. Speech Synth*. Cham, Switzerland: Springer, 1997, pp. 313–323.

[58] S. Flores Solórzano and R. Coto-Solano, "Comparison of two forced alignments systems for aligning Bribri speech," *CLEI Electron. J.*, vol. 20, no. 1, pp. 2:1–2:13, 2017

[59] S. Young et al., *The HTK Book*, vol. 3. Cambridge, U.K.: Cambridge Univ. Press, Engineering Department, 2002, p. 175.

AMMAR MOHAMMED ALI ALQADASI received the B.S. degree in information technology from Taiz University, Taiz, Yemen, in 2008, the M.S. degree in software engineering and intelligence from UTeM, Melaka, Malaysia, in 2012, and the Ph.D. degree in computer science form UTM, Johor, Malaysia, in 2021. His research interests include speech processing, image processing, and intelligent computing.

AKRAM M. ZEKI (Senior Member, IEEE) received the B.Sc. degree from the University of Jordan, Amman, Jordan, in 2000, the master's degree in computer graphics from the Faculty of Computer Science and Information Technology, University Putra Malaysia, in 2004, and the Ph.D. degree from the Faculty of Computer Science and Information System, University Technology Malaysia, in 2009. He did his master's research with the Kulliyyah of Information and Communication Technology, International Islamic University Malaysia (IIUM), in 2009, where he is currently an Associate Professor and the Deputy Director of the Centre for Islamisation (CENTRIS). Apart from supervising master's and Ph.D. students, he is also involved in research. He has published seven books, 20 articles in journals, 45 book chapters, and more than 75 conference papers. His research interests include watermarking, steganography, information security, image processing and multimedia, IT, and Islam and develop Islamic applications. Besides that, he has received awards at both university and national levels.

MOHD SHAHRIZAL SUNAR (Member, IEEE) received the B.Sc. degree in computer science majoring in computer graphics from Universiti Teknologi Malaysia, the M.Sc. degree in computer graphics and virtual environment from the University of Hull, U.K., and the Ph.D. degree from the National University of Malaysia, in 2008. He is currently the Director of the Institute of Human Centered Engineering (iHumEn) and the founding Director of the Media and Game Innovation Centre of Excellence (MaGICX), Universiti Teknologi Malaysia. His major field of study is real-time and interactive computer graphics and virtual environment. He received scholarships from Sultan Iskandar Johor Foundation for both his postgraduate study.

MD. SAH BIN HJ SALAM received the B.S. degree in computer science from the University of Pittsburgh, PA, USA, and the M.S. and Ph.D. degrees in speech processing and AI from Universiti Teknologi Malaysia. He is currently the Head of the Research Vicubelab Group and a Senior Lecturer with the Faculty of Computing, Universiti Teknologi Malaysia. His research interests include speech processing and synthesizing, image processing, intelligent computing, and computer graphics.

RAWAD ABDULGHAFOR received the bachelor's and master's degrees in computer systems engineering and informatics from Saint Petersburg Electrotechnical University "LETI," Russia, in 2004 and 2007, respectively, and the Ph.D. degree in information technology from International Islamic University Malaysia, Kuala Lumpur, Malaysia, in 2017. He was a Research Assistant with the Faculty of Information and Communication Technology, International Islamic University Malaysia, from 2014 to 2017. He has been an Assistant Professor with the Faculty of Information and Communication Technology, International Islamic University Malaysia, since 2018. He was a Postdoctoral Fellow with the Malaysia-Japan International Institute of Technology, Universiti Teknologi Malaysia, Kuala Lumpur, from 2018 to 2019. His research interests include consensus models for multi-agent systems, computation theory, wireless communications, and machine learning.

NASHWAN ABDO KHALED received the degree in Qira'at from Taibah University, Saudi Arabia, and the master's degree in interpretation of the Quran (Tafsir Maqasidi) and minoring in the area of (Maqasid Al-Quran) and the Ph.D. degree in Quran and Sunnah studies from International Islamic University Malaysia (IIUM). He has been an Assistant Professor with the Department of Quran and Sunnah Studies, since November 2017. He is currently the Founder of ITQAN Foundation, a Quranic organization that caters to social and public needs. He has published more than seven books chapter in books and 20 refereed articles in international journals in both Arabic and English. His research interests include serve the holy Quran in every aspect of life—academic, social, and public. He has been the Chairperson of the Postgraduate Committee with the Department of Quran and Sunnah Studies, since 2019. He is also a Board Member of *Al-Burhān Journal of Qur'an and Sunnah Studies* with IIUM.