

# UMT Artificial Intelligence Review (UMT-AIR)

Volume 3 Issue 2, Fall 2023

ISSN(P): 2791-1276, ISSN(E): 2791-1268

Homepage: <https://journals.umt.edu.pk/index.php/UMT-AIR>



Article QR



**Title:** Recitation of The Holy Quran Verses Recognition System Based on Speech Recognition Techniques

**Author (s):** Muhammad Rehan Afzal<sup>1</sup>, Aqib Ali<sup>2</sup>, Wali Khan Mashwani<sup>3</sup>, Sania Anam<sup>4</sup>, Muhammad Zubair<sup>2</sup>, Laraib Qammar<sup>5</sup>

**Affiliation (s):** <sup>1</sup>The Islamia University of Bahawalpur, Bahawalpur, Pakistan

<sup>2</sup>Concordia College Bahawalpur, Pakistan

<sup>3</sup>Kohat University of Science & Technology, Kohat, Pakistan

<sup>4</sup>Govt Associate College for Women Ahmedpur East, Bahawalpur, Pakistan


<sup>5</sup>Bahauddin Zakariya University, Multan, Pakistan

**DOI:** <https://doi.org/10.32350/umt-air.32.01>

**History:** Received: December 22, 2022, Revised: April 4, 2023, Accepted: May 28, 2023, Published: December 2, 2023

**Citation:** M. R. Afzal, A. Ali, W. K. Mashwani, S. Anam, M. Zubair, and L. Qammar, "Recitation of the Holy Quran verses recognition system based on speech recognition techniques," *UMT Artif. Intell. Rev.*, vol. 3, no. 2, pp. 00–00, Dec. 2023, doi: <https://doi.org/10.32350/umt-air.32.01>

**Copyright:** © The Authors

**Licensing:**  This article is open access and is distributed under the terms of [Creative Commons Attribution 4.0 International License](https://creativecommons.org/licenses/by/4.0/)

**Conflict of Interest:** Author(s) declared no conflict of interest



A publication of

Department of Information System, Dr. Hasan Murad School of Management  
University of Management and Technology, Lahore, Pakistan

# Recitation of The Holy Quran Verses Recognition System Based on Speech Recognition Techniques

Muhammad Rehan Afzal<sup>1</sup>, Aqib Ali<sup>2\*</sup>, Wali Khan Mashwani<sup>3</sup>, Sania Anam<sup>4</sup>, Muhammad Zubair<sup>2</sup>, and Laraib Qammar<sup>5</sup>

<sup>1</sup>Department of Computer Science, The Islamia University of Bahawalpur, Bahawalpur, Pakistan

<sup>2</sup>Department of Computer Science, Concordia College Bahawalpur, Pakistan

<sup>3</sup>Institute of Numerical Sciences, Kohat University of Science & Technology, Kohat, Pakistan

<sup>4</sup>Department of Computer Science, Govt Associate College for Women Ahmedpur East, Bahawalpur, Pakistan

<sup>5</sup>Department of Computer Science, Bahauddin Zakariya University, Multan, Pakistan

**ABSTRACT** Arabic is the language in which the Holy Quran was revealed to Mohammed (S.A.W). Muslims claim that the Holy Quran has not been tampered with since it has been preserved. The Arabic Quran should be read exactly as it has been written. With the flourishing of Islam and the appearance of faults in Quran's recitation, the experts created Tajweed to preserve Allah's revelation. The Holy Quran's authenticity and purity must be protected from erasure or contamination. The current study examined speech recognition techniques used in the Quran's recitation along with their strengths and faults. Moreover, it also examined the Quranic text verification paradigm. The development of a computer-aided system, to automatically learn the Holy Quran's recitation, is a practical learning technique. Computer-aided Programming Language (CAPL) has gained popularity in recent years. Moreover, numerous researches have been conducted so far to improve these methods, especially in second-language instruction. Computer technologies can help language teachers with pronunciation and accent reduction. The computers play an essential role in automated tutoring. With the help of computer, words can be learned at home. CAPL's strict application is to automate the Holy Quran's recitation unlike a language-learning exercise, where many pronunciations may be appropriate. There is minimal opportunity for variation while reciting the Holy Quran in Arabic language. The current study presented a concept for Quran's recitation verification system along with an overview of Quran's voice recognition techniques.

**INDEX TERMS** Holy Quran, machine learning, recitation, speech recognition

## I. INTRODUCTION

Muslims believe that Gabriel brought the Holy Quran to Muhammad (S.A.W.), revealed by Allah. We, Muslims, believe that the Holy Quran is a divine revelation. We must obey Allah's rules to gain His

pleasure and place in heaven. The Holy Quran defines the purpose of our life and imparts moral, social, and spiritual values as well. The Quran portrays Muhammad (S.A.W.) as the ideal leader. The Quran's prophecies affect its believers' lives.

---

\*Corresponding Author: [aqibcsit@gmail.com](mailto:aqibcsit@gmail.com)

Muslims recite the Holy Quran during five daily prayers and on other occasions as well. The Holy Quran has 323,015 letters, 77,439 words, over 6000 verses, and 114 chapters [1]. Hafs from A'asim, Warsh, Qalun, and others have narrated the Holy Quran. Every narrative is location-specific. The Holy Quran is unlike any other Arabic script. The Tajweed regulations, which govern how the Quran is spoken, are highly severe [2]. Each narrative of the Holy Quran has Tajweed regulations, such as Warsh and Hafs from A'asim. However, some principles apply to all narrations, such as the rule of mandatory extension. The primary goal of Tajweed is to modify the Quran's recitation procedure, so that all the Muslims may recite the Quran correctly [3].

The speaker-independent system is trained to detect speech regardless of who talks. It allows us to get numerous patterns of speech inflections (pitch, tone, voice rise, and fall) and enunciation of targeted words with high accuracy within processing restrictions [4]. Siri, Google Assistant, and Samsung's Bixby use speaker-independent method. Some of this work was done for dictation tools, such as I.B.M. through voice Arabic [5]. BBN Tides-On-Tap technology is a newly introduced Arabic A.S.R. application. The BBN Arabic broadcast news recognizer makes 15% of word errors (WER). The first large-scale attempt to construct recognizers for conversational (dialectal) Arabic rather than M.S.A. was in 1996 and 1997 as part of NIST Call Home benchmark tests. These tests compared phone voice recognition algorithms in several languages including Egyptian Colloquial Arabic. BBN's systems performed best in these tests, with error rates between 71.1% and 61.66%. Speech processing is more complex as compared to text classification and image recognition. Speech recognition combines

signal processing, phonetics, linguistics, and machine learning [6]. Success requires thorough understanding. The Arabic language needs speech-to-text converters. Arabic voice recognition research involves the creation of a small internal corpus. For years, speech-to-text projects have employed Cambridge's Hidden Markov Model (HMM's) Toolkit and C.M.U.'s Sphinx (H.T.K.). Both engines use HMMs. Building a new speech recognition engine from scratch is a complicated task which requires professional programming skills. Most academics utilize free research A.S.R. engines, such as H.T.K. and Sphinx. SVMs, ANNs, and k-nearest neighbours (KNN) are alternative (or combination) identification algorithms. With deep learning, deep neural networks have emerged as an exciting voice recognition topic [7].

Signal processing was utilized to identify and fix the pronunciation difficulties. MFCC coefficient, signal energy, delta MFCC, and delta-delta MFCC are extracted to eliminate voices. Voice signal processing is a time-and frequency-dependent discrete signal. The vocal tract is a key to voice production. Firmness, tension, vocal cord length, and airflow in the glottis tend to affect the vocal cord frequency. This frequency component consists of fundamental harmonics. Unvoiced means without vocal cord vibrations [8]. The Holy Quran is regarded as a sacred text. Every Muslim is obligated to believe that the text of the Holy Quran has remained unchanged since it was first revealed to the Prophet Muhammad (S.A.W). As a direct result of the expansion of multimedia on social networks, several individuals who pursue various goals have availed the chance to disseminate the Holy Quran. There are both, positive and negative reasons for doing anything.

Although, some people may have questionable objectives, such as distorting and tampering with the Holy Quran. The positive side is that it may be used to teach an authentic message to the people around the globe. It may appear from the audio clips that the Holy Quran does not have a specific identifying technique [9].

## II. LITERATURE REVIEW

The references [10] stated that loudspeakers impair ASR performance. EEG may help the ASR systems to avoid noise-induced performance degradation. Tacotron's natural-sounding multi-speaker synthetic speech has prompted interest in replacing pricey, painstakingly transcribed human audio, used to train speech recognizers. The references [11] proposed pre-trained unsupervised speech utilizing raw audio representation system. The WER of a strong character-based log-Mel filter bank baseline was decreased by up to 36% in the current study on the WSJ. On nov92, the approach delivered 2.43% WER. This outperforms deep speech 2, the best-documented character-based system. The references [12] improved Libri Speech using unlabelled Libri-Light audio. On the Libri Speech test/test-other sets, WERs of 1.4%/2.6% were reached as compared to 1.7%/3.3%.

The references [13] developed hybrid speech recognition by using transformer-based acoustic models (AMs). This work analysed the positional embedding and iterated profound transformer loss modelling methodologies. The early transformer model streaming research was discussed by using constrained proper context. In small voice recognition datasets, references [14] found Speech Transformer, a no-recurrence encoder-decoder architecture, promising. The current study optimized the Speech Transformer for a

large-scale Mandarin Chinese speech recognition as a challenge. The frame rate was reduced to optimize processing and model performance—scheduled sampling and concentrated loss lower character mistake rates (CER). On four test sets, the recommended modifications increase the CER by 10.8% to 26.1% on 8,000-hour work. The final improved Speech-Transformer reduces CER by 12.2%-19.1% without explicit language models as compared to a powerful hybrid TDNN-LSTM system trained with LF-MMI criteria and decoded with a big 4-gram LM. End-to-end (E2E) models replaced hybrid models after introducing automatic speech recognition. E2E techniques include RNN-T, Transformer-AED, and RNN attention-based encoder-decoder.

The references [15] recommended E2E automated speech recognition and pseudo-labelling (ASR). With the advancement of acoustic model, the examination of Iterative Pseudo-labelling (IPL), a semi-supervised pseudo-labelling method, would be conducted. The IPL refines a model by using labelled and unlabelled data to study the language model by decoding and data augmentation. Moreover, it also examines how language models affect IPL's text consumption. Low-resourced and semi-supervised ASR research should be supported. The references [16] proposed multilingual E2E models for global ASR coverage. Eliminating language-specific acoustic, pronunciation, and language models improves training and service over monolingual systems. The current research presents an E2E multilingual system for low-latency interactive applications and real-world data imbalance. The best model uses language vector conditioning and language-specific adaptor layer training.

The references [17] reviewed the Automatic Tajweed verse recitation rules

engine. This study combined speech, semantics, and stemmers with questions and vice versa. It also evaluated retrieval and relevance. Polysemy retrieved irrelevant documents. CLIR requires SSSQ, according to trials. The analysis of the Holy Quran's document outcomes with stemmers was superior to voice. The references [18] represents Automatic Quran's pronunciation error detection. The current study found mispronounced emphatic and non-emphatic characters to create algorithms in order to extract the distinguishing feature from phrases, containing the letters of interest.

#### A. ASR Systems of Al-Quran Recitation

Numerous scholars have suggested Al-Quran's recitation verification methods to meet its challenges. They do not verify Al-Quran's recitation for Arabs and non-Arabs, according to all Tajweed requirements, which is a crucial difficulty for reciting the Quran accurately [19]. Traditional face-to-face Al Quran recitation teaching is embedded with many issues.

- A teacher with many students cannot pay attention to each one of them.
- Teachers are not always available.
- Accessing Tajweed literature is difficult and time-consuming.
- Same teaching style for all pupils, regardless of comprehension levels.
- Some learners are visual and some are audio.

Due to the above-mentioned challenges, ASR researchers have created automated methods. These methods allow the users to evaluate their recitation quickly and easily anywhere. This technique covers the first level of Tajweed with Rewaya Hafs from "Aasem". Students and lecturers tested this system [20]. The current research considered the unique approaches to recite the AL-Quran with Tajweed standards.

Moreover, it also aimed to build a revolutionary strategy that uses voice recognition algorithms to automatically find and delimit the verses in audio recitations, regardless of the reciter. The study also employed the HMM-based Sphinx Framework as a research platform and the Sphinx Train to generate the acoustic models. MFCC, Sphinx, and acoustic models were used to extract system characteristics for recognition [21]. HAFSS is a CAPL system for non-native teaching speakers of Arabic pronunciation and Al-Quran recitation. This technology offers feedback on recitation problems and how to rectify them.

### III. PROPOSED METHODOLOGY

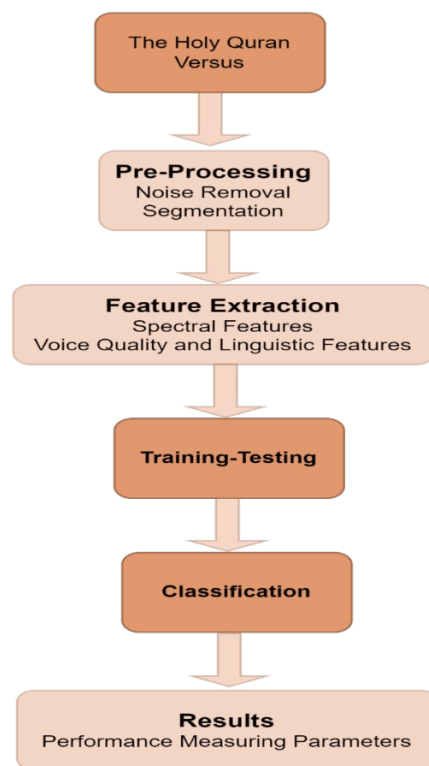


FIGURE 1. Architecture of voice recognition system

The current study attempted to develop a system to recognize the verses of Holy Quran by using different reciters' recitation data collection. It requires a transcription file, a phonetic dictionary, a set of phonemes, and the audio file. The collection of the Holy Quran verses recitation will be of 5 different reciters. The process is described in six basic steps as shown in Figure 1.

- The first step is the collection of data for the recitation of Holy Quran's versus. In this step, different reciters' recitations are collected in audio format.
- The second step is data processing. In this step, the data-cleaning process is applied along with the extraction of feature according to its patterns. The qualitative nature of data would be examined along with the extraction of useful information to help the current study.
- The third step is the extraction of parameters which involves extracting the parameters of the audio versus recitation dataset.
- The fourth step is feature extraction. In this step, different speech features are extracted. For instance, removing speaker-dependent characteristics, identification of Arabic language phonemes, prosodic features, voice quality, and linguistic features, and get spectral features of the frame using the (CMU) Sphinx tool.
- The fifth step is classification. After following the above steps, the dataset is prepared. Now, different classification algorithms would be applied, such as the DWT algorithm, on the dataset, to get the classification results by using HTK and 16 (GMMs). The related dataset would be classified according to its pattern.

- In the sixth and final step, the testing dataset would be compared with the training dataset to identify and recognize the accuracy of voice. By performing different experiments and evaluation processing, results in getting all Tajweed rules are divided into two classes (matched or Not Matched).

**A. Data Collection**

Speech processing needs a consistent dataset. No standard mispronunciation dataset is available. The lack of a standard voice corpus for mispronunciation recognition may be attributed to language dependence and subjective labelling. The current research's dataset generation involved three steps:

- Designing and collecting datasets.
- Annotating and tagging datasets.
- Annotation and labelling are considered as bottlenecks in mispronunciation detection dataset recording. Linguists label the things and transcribe them. This transcription matches the pronunciation dictionary's order. Moreover, labelling also involves grading each phoneme's quality.

Khaa	Haa	Jim	Thaa	Taa	Baa	Alif
خ	ح	ج	ث	ت	ب	ا
Saad	Shiin	Siin	Zay	Raa	Dhaal	Daal
ص	ش	س	ز	ر	ذ	د
Qaaf	Faa	Ghayn	Ayn	Zaa	Taa	Daad
ق	ف	غ	ع	ظ	ط	ض
Yaa	Waaw	Haa	Nun	Miim	Laam	Kaaf
ي	و	ه	ن	م	ل	ك

**FIGURE 2.** Arabic alphabetic representation

## 1). DATASET DESIGN AND COLLECTION

**SUBJECT.** The speakers belonged to Pakistani descent and from various demographic groups as shown Table I. Pakistan is a demographically varied country, with individuals from various cities speaking various mother languages.

TABLE I  
DATASET USED IN THIS  
EXPERIMENT

Mother Tongue	No of Males	No of Females
Urdu	5	5
Punjabi	5	5
Pashto	5	5

TABLE II  
NUMBER OF PHONEMES WITH  
LABELS

Training		Testing	
Native	Non-Native	Native	Non-Native
10	10	10	10

## 2) DATASET ANNOTATION AND LABELLING

Linguists' subjectivity limits the dataset labelling. Language experts categorize the datasets differently. Firstly, linguistic experts discuss the topic. Afterwards, the language professionals address labelling

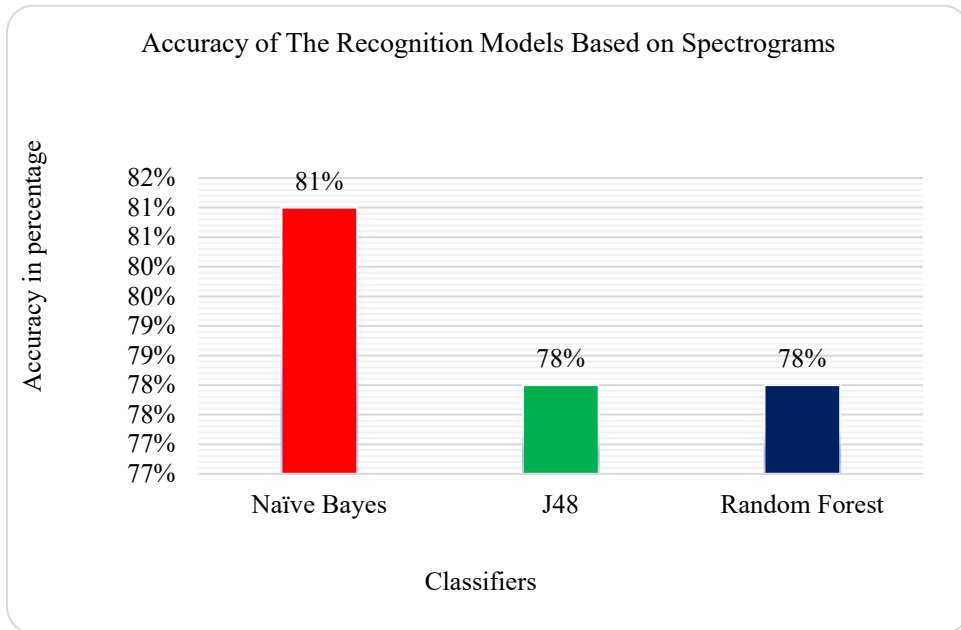
instructions and their benefits. Forced alignment separates phonemes from the signal. Five Arabic experts from Pakistan teaching Tajweed tag the dataset. Language experts categorize the Arabic phonemes separately as shown in Table II.

## B. AUDIO REPRESENTATION

Tajweed is a requirement for Muslims to both recite and hear the Holy Quran. The current study identified the Holy Quran reciters who employ machine learning [22]. Twelve Qaris recited ten more Surahs from the database of the current study. These 12 Qaris symbolized the 12-class difficulty. Audio is the first frequency domain to be examined before being processed into a Spectrogram. MFCC and pitch are utilized initially as model learning properties. The characteristics are extracted from audio as visuals using auto-correlograms [23]. Considering their cutting-edge performance, these classifiers were selected. Moreover, an accuracy of 88% was achieved in recognizing Qari from Quranic recitations by using Nave Bayes and Random Forest. It demonstrated that Qari can be identified.

## 1). RECOGNITION MODELS AND FEATURES

The classifiers and feature extraction serve the primary purpose of experimentation and analysis. Figure 3 and Table III describe the accuracy percentage of different classifiers during speech recognition.



**FIGURE 3.** Spectrogram features performance analysis

TABLE III  
SPECTROGRAM FEATURES  
PERFORMANCE ANALYSIS

Classifier	Accuracy
Naïve Bayes	81%
J48	78%
Random Forest	78%

**C. CMU SPHINX TOOL**

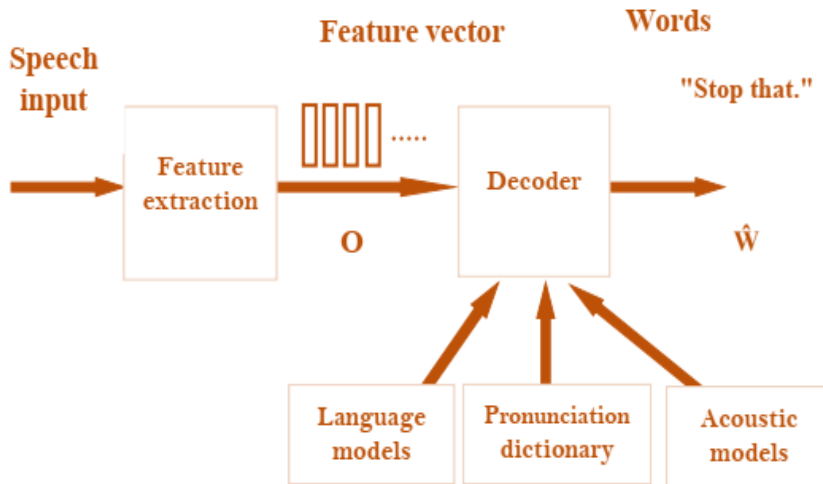
CMU Sphinx contains numerous libraries and training tools. Kai-Fu Lee and his colleagues created Sphinx 1 and currently use Sphinx 4. CMU Sphinx is the first large-vocabulary continuous speech recognizer (LVCSR) [24]. It is a resilient and adaptable open-source initiative for voice recognition research. Carnegie Mellon University's Sphinx group teamed with MERL, Sun Microsystems, and HP's

Cambridge Research Lab in 1987 to create this open-source speech recognizer (HP). UC and MIT funded CMU Sphinx (MIT). High-performance speaker-independent English ASR [25] system uses three states of discrete HMM, 256 vocabularies, and has an 89% accurate recognition rate. Sphinx 2 employs a C-written five-state semi-continuous HMM with probability density functions. Wall Street Journal speech database improved Sphinx 2 accuracy to 90% [26].

1) STRUCTURE OF CMU SPHINX

The CMU Sphinx recognizer uses hidden HMMs for voice recognition [27]. This channel includes speaker's vocal equipment which creates the speech waveform and the speech recognizer X's signal processor. Figure 4 shows major components and procedures of the CMU Sphinx recognizer as shown in Figure 4.





**FIGURE 4.** Architecture of CMU sphinx recognizer

#### ***D. PRE-PROCESSING WITH CMU SPHINX TOOL***

##### ***1) PREPARATION DATASET WITH TOOL***

The first step is to prepare the corpus which is a collection of unit sounds described by specific words in the lexicon (Training Acoustic Model for CMU Sphinx, n.d.). All the ten Arabic numerals were used to generate the corpus (zero to nine). Thirty native Arabic speakers were instructed to repeat all numerals ten times. Resultantly, every digit produced by each speaker was repeated ten times in the database as shown in Table IV.

#### ***E. PROCESS OF EXTRACTING FEATURES***

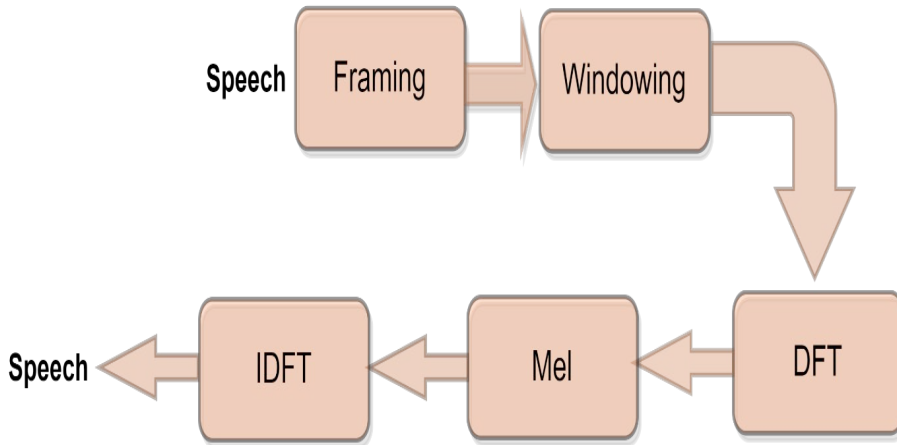
In order to identify between a wide variety of different words, feature extraction extracted crucial characteristics, specific to each word from an audio signal. This is because, as compared to other feature

extraction methods, MFCC may produce results with greater accuracy while requiring less computing complexity [28].

TABLE IV

PARAMETERS OF RECORDING SYSTEM

Parameter	Value
Sampling Rate	17 kHz, 17 bits
Wave format	Mono, wav
Corpus	Isolated 10 Arabic digits (zero to nine)
Arabic Native speakers	20 males
Non-Native Arabic Speakers	20 Males
Repetition	10 Times
Window type and size	Hamming, 256



**FIGURE 5.** MFCC Processes

MFCC is used in speech processing to extract features as shown in Figure 5. This approach introduced Mel scale cepstrum coefficients. The Mel scale maps linear frequency to human auditory perception. The speech waveform is sliced to remove silent or acoustic interference at the start or end of the sound stream. This enables Fourier transformation. The hamming window process minimizes and eliminates the signal frame discontinuities [29]. It is the most used MFCC approach for decreasing the signal discontinuities by zeroing off each frame. It compares Hz and Mel frequencies. Mel-scaling requires several triangular filter banks. Therefore, MFCC generates one for each frequency band. Inverse Discrete Fourier Transform MFCC feature extraction ends with inverse DFT. This step creates voice-feature vectors. Feature vectors are retrievable. The next phase's input is this feature vector. Both, speaker-independent and speaker recognition use MFCC. In the current study, the incoming speech signal was sampled at 8000 Hz by using the Nyquist

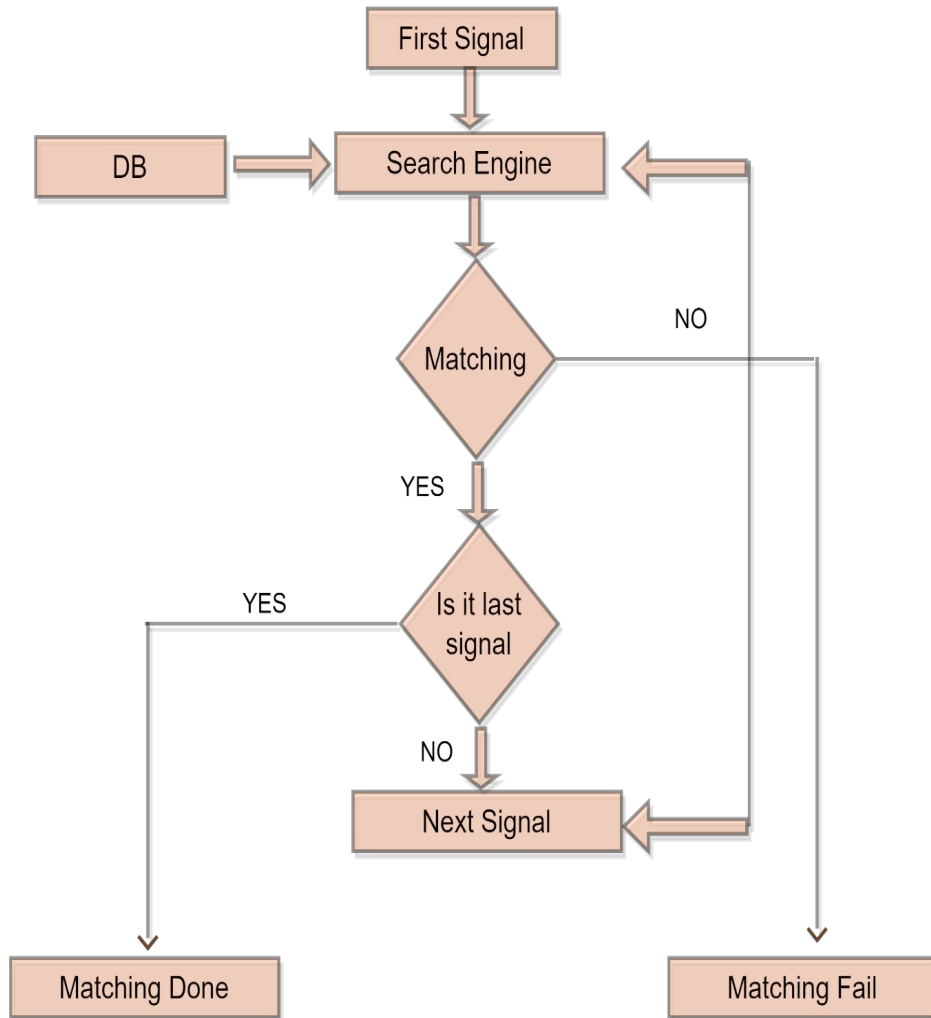
technique and split into 40-ms frames with a 20-ms overlap, resulting in a 50% overlap. The number of recited words determines frame separation. The signal is converted to the frequency domain by using  $N=1024$  DFT. About 24-triangle filters handle the signal. After filtering, 14-MFCC DCT is conducted [30].

#### 1) MODELLING AND STORING

In this stage, a model representing a recitation feature vector is created by using the output from the MFCC process. Any verse submitted by the user is checked against verses preserved in the different Hafizes of the Quran. The user is alerted and the verses that do not line up are marked as mistakes [31].

#### 2) SEARCHING AND MATCHING

In this stage, audio signals are gathered, and a database containing all Quranic verses saved as MFCC factors is searched for matching audio signals [32]. This phase is further elaborated in Figure 6.



**FIGURE 6.** Block diagram of search and matching step

#### IV. EXPERIMENTS AND RESULTS

##### • Comparative Analysis

In the current study, different pattern recognition classifiers implemented the dataset. This comparative analysis was performed and different performance calculating parameters learned classification accuracy results.

##### • TP-Rate

The true positive rate, also known as sensitivity or recall, is a measurement used in machine learning. It helps to determine the proportion of real positives that are accurately identified. True Positive Rate can be calculated as:

$$\text{TP-Rate} = \text{TP} / (\text{TP} + \text{FN})$$

- **TN-Rate**

True Negative Rate can be calculated as

$$TN\text{-Rate} = TN / (TN + FP)$$

- **FP-Rate**

False Positive Rate can be calculated as

$$FP\text{-Rate} = FP / (FP + TN)$$

- **FN-Rate**

False Negative Rate can be calculated by

$$FN\text{-Rate} = 1 - TP\text{-Rate}$$

- **Precision**

$$PRECISION = TP / (TP + FP)$$

- **Recall**

$$RECALL = TP / (TP + FN)$$

- **F-Measure**

$$F\text{-Measure} = 2 * Precision * Recall / (Precision + Recall)$$

### A. RESULTS USING SFS

K Nearest Neighbours (KNN) was utilized as a classifier with the updated version of SFS. The classifier was trained by using a unique feature set for each phoneme due to the fact that each phoneme could be discriminated against using a unique collection of acoustic data.

TABLE V

RESULTS OF CLASSIFICATION USING THE K-NN CLASSIFIER AND SFS FOR k=9

Results After Classification		
No. of Feature Set	Average No. of selected Features	Average Accuracy with SFS
289	102	92.15%

The Table above compares the proposed method with current systems. These systems have statistical CALL features. CALL for English has been a standard since its creation. Each phoneme's GOP score was utilized to determine the accuracy. This method uses statistical characteristics to evaluate the pronunciation. This method was 80-92% accurate. The recommended approach detected mispronunciations with 92.15% accuracy while using only auditory criteria. This was done without an ASR or a well-known mathematical model. Statistical CALL was established to identify mispronunciations. Only four mispronunciation classifiers were created. One classifier trained on auditory features performed the best. The method was 81-88% accurate. The recommended system outperformed the present technique. The recommended approach features 28 Arabic phonemes, yet it only caught one pronunciation problem. A feedback-based CALL system teaches non-native Dutch speakers' accurate pronunciation. GOP pronunciation ratings were 86% accurate. A separate classifier addressed each pronunciation mistake. This strategy addressed pronunciation issues, speaker variances, and phoneme duration mistakes. Robust HMM is used to discover recitation mistakes, while only 52% of the mistakes were caught. The proposed system solved these flaws and outperformed the present one. A second CALL approach identifies six mispronounced Arabic phonemes by using the statistical data. GOP scores were used to identify mispronunciations. The class-based average was calculated to be 92.95%. Similar results were generated by using acoustic qualities. The suggested system uses 28 phonemes instead of six already used. The results showed that the recommended method might be able to deliver the best results. A primary classifier is employed to obtain comparable results in

order to determine the phoneme characteristics. These findings suggest that even a simple classifier may produce excellent results if feature selection is made appropriately and discriminative features are used for training.

TABLE VI  
COMPARISON OF SUGGESTED METHOD WITH TRADITIONAL ARABIC CAPT SYSTEMS

Mispronunciation Detection Systems for Quran recitation						
Techniques	Proposed Techniques	[33]	[34]	[35]	[36]	[37]
Avg. accuracy	92.15%	52.2%	81-88%	92.95%	80-92%	86%

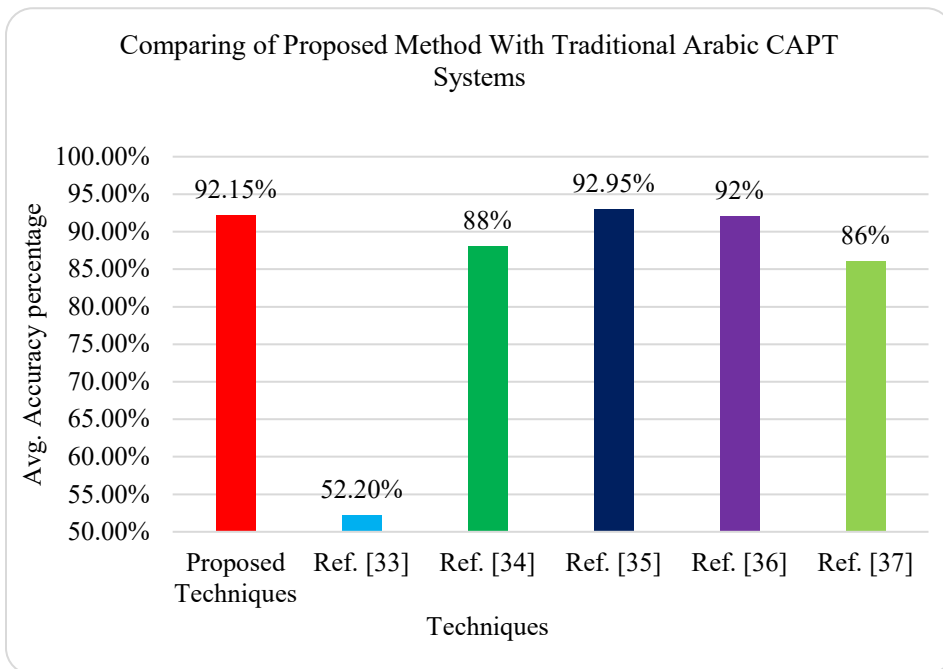


FIGURE 7. Comparison of suggested method with traditional Arabic CAPT systems

This work evaluated the effectiveness of extensive acoustic-phonetic features for mispronunciation-detecting systems. Sequential Forward Selection (SFS) was used to choose discriminative qualities for each Arabic phoneme. The best discriminate collection of speech acoustic features was determined by using SFS. SFS

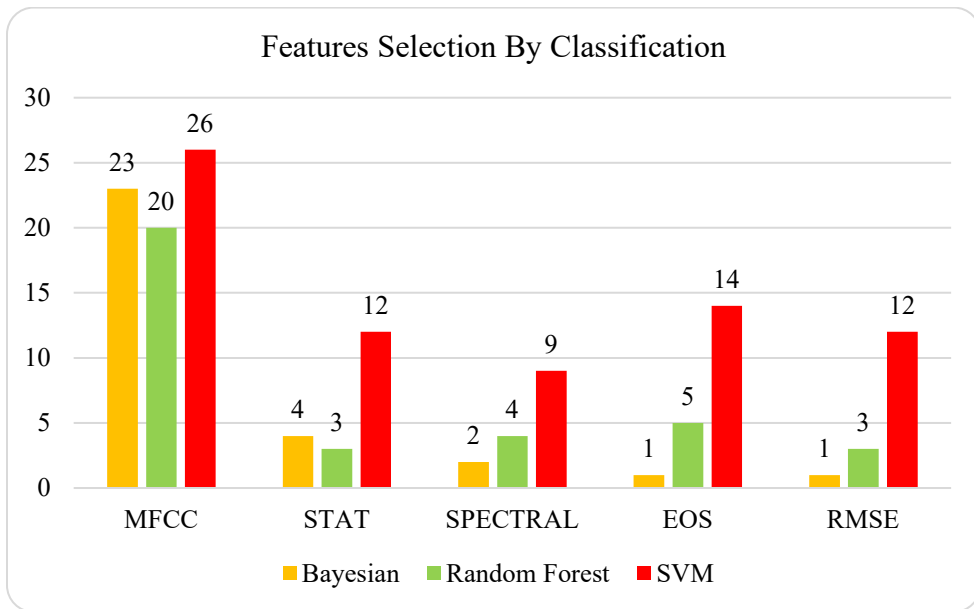
procedures cease when the system's accuracy declines, making it difficult to analyse every feature. All the research characteristics are updated and SFS-tested. To fix this, more feature selection techniques would be attempted. Mispronunciation detection, using the K-NN classifier, shows that a primary

classifier may produce 58 correct results when combining most discriminative variables.

**B. RESULTS OBTAINED BY USING SFFS TECHNIQUE**

These criteria were evaluated for each Arabic consonant's pronunciation. All accuracy numbers were rounded. Each Arabic consonant's feature was compared to each classifier. These classifiers were running independently with default parameters to ensure phoneme discrimination. Top-performing Arabic consonant characteristics are underlined in the related tables.

Other acoustic features show promise beyond MFCCs and their first and second derivatives. These features include RMSE, Statistical Features, and EOS (RMSE). In some instances, these traits outperform MFCCs. Pitch and ZCR are often used to classify speech. EOS performed somewhat better than spectral characteristics, while both contributed to mispronunciation detection. Only statistical features regularly outperformed MFCCs. MFCCs surpassed statistical features by a large margin. The total signal effect led to excellent statistical results. Pitch, low energy, and ZCR could not assist to detect mispronunciations. While, these qualities performed poorly overall, certain consonants helped (Figure 8).

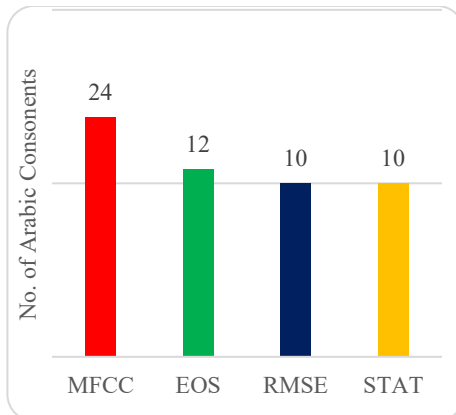


**FIGURE 8.** Comparison of the Top Three Characteristics Chosen by Classifiers that Perform the Best

Bagged SVM obtained 94.9% accuracy and 0.067 MAE. It outperformed the other two classifiers. Bagged SVM was selected to discover the relevant characteristics. Each

top three consonant forming features was counted. MFCCs were voted as the top feature 27 times and Entropy of Spectrum 14 times. The top three features had 12

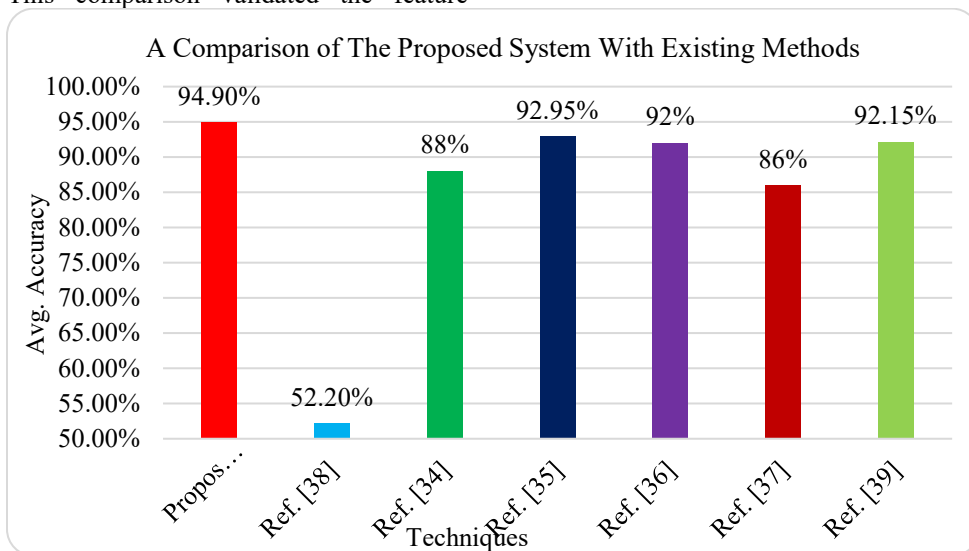
RMSE and Statistical occurrences. MFCCs excelled all consonants.



**FIGURE 9.** Top 3 Highest-performance features contribution using bagged SVM

This system allowed for all features despite greedy algorithms' drawbacks. It evaluates every aspect to determine the best acoustic feature combination. Without this change, SFFS would cease when its accuracy drops. The top three hand-picked phoneme characteristics were compared to SFFS. This comparison validated the feature

choices. Both hand-picked and SFFS selected the same features. SFFS picked over three qualities for most phonemes. SFFS's feature vector includes top three manual-testing features. Sometimes, feature selection chooses unnecessary features. Both techniques' hand-picked features generated better classification results than SFFS. For each phoneme, sonic features were repeated. This category includes MFCCs and first and second derivatives. MFCCs and their first and second variations passed 26 14 12 12 5 0 10 15 20 25 30. The number of Arabic consonants contributed to top 3 features of 62 pronunciation teaching systems. MFCCs alone was not recommended. MFCCs are generic, however, other acoustic-phonetic features are pronunciation-specific. The results showed a high association between hand-selected and SFFS' attributes. Most consonants had good correlation coefficients. SFFS did not choose many consonant traits. Consonants had a low correlation coefficient. The average correlation was 0.82.



**FIGURE 10.** A comparison of the proposed system with the existing methods

The present method has better acoustic phonetic qualities than the previous one. Other noteworthy systems had accuracy ratings of 86%, 80-92%, and 81-88%. These strategies used confidence ratings. Another technique uses feature selection

and a simple classifier to detect mispronunciations. About 92.15% were accurate. The present method's accuracy surpassed all others. The results showed that the recommended technique chose the best Arabic consonant qualities.

TABLE VII

A COMPARISON OF THE PROPOSED SYSTEM WITH THE EXISTING METHODS

Mispronunciation Detection Systems for Quran recitation							
Techniques	Proposed Techniques	[38]	[34]	[35]	[36]	[37]	[39]
Average Accuracy	94.9%	52.2%	81-88%	92.95%	80-92%	86%	92.15%

**V. CONCLUSION AND FUTURE WORK**

We are developed an acoustic-phonetic CALL system for The Holy Quran vs. recitation recognition. A standard dataset, to recognize the Arabic mispronunciations, was also absent. According to the current study, no computer-aided language learning dataset is publicly available. The Holy Quran vs. recital and recitation datasets were obtained for the current study. This dataset comprises Pakistanis from diverse socioeconomic backgrounds. The employed system is the most advanced CALL system, primarily based on ASR. It cannot detect pronunciation problems in a learner's voice. Mispronunciation detection techniques, based on confidence measures, can only transmit a recognizer's level of accuracy. Automated mispronunciation feature selection is a research challenge. Each mispronunciation needs a classification. Two parts have explored these concerns in the current study. The first half of the study discussed how a basic feature selection technique may improve

mispronunciation detection systems by considering it a classification problem.

SFS and SFFS are utilized for feature selection and recognition. About 96 automatically selected acoustic-phonetic components increased the results. FLDA and PCA reduce dimensionality by determining the discriminative audio characteristics (PCA). Subjective analysis displays how closely outcomes match personal views. The current study's dataset collection may have altered the experiment outcomes and raised validity concerns, such as instrumentality and selection bias. Soundproof rooms are used to collect acoustic data. The study's speakers reflect a wide range of demographics, making it impractical to get them all in a soundproof room. The entire dataset was recorded in an open office, introducing noise to the audio. All of the experiment's accuracy may have decreased. Gender and age may have altered the trial recommendations. Gender- and age-specific speech processing systems perform better. This field has several unexplored research paths. The automatic identification of pronunciation errors



requires more complex methods. Dimensionality-reduced feature selection was offered which improves mispronunciation-detection systems. All language-based segmentation methods need 97 transcripts. Language transcripts should be independent of segmentation. A language-independent mispronunciation detection system is needed immediately. System design requires data standardization. Dimensionality reduction and categorization can detect mispronunciation in any language. Super-vectors are calculated by utilizing a language's acoustic-phonetic properties. FLDA and PCA may leverage language-specific features to detect misspellings. Lastly, categorization is performed. Contrasts may differ in subsequent rounds. A new, effective algorithm for sparse acoustic-phonetic qualities can be developed.

## REFERENCES

- [1] A. A. Abdullah, M. D. Awang, and N. Abdullah, "Islamic tourism: The characteristics, concept and principles," *KnE Soc. Sci.*, Jul. 2020, pp. 196–215, doi: <https://doi.org/10.18502/kssv49.7326>
- [2] R. Hakim, M. Ritonga, K. Khodijah, Z. Zulmuqim, R. Remiswal, and A. R. Jamalyar, "Learning strategies for reading and writing the quran: improving student competence as preservice teachers at the faculty of tarbiyah and teacher training," *Educ. Res. Int.*, vol. 2022, Art. no. 3464265, June 2022, doi: <https://doi.org/10.1155/2022/3464265>
- [3] I. K. Tantawi, M. A. M. Abushariah, and B. H. Hammo, "A deep learning approach for automatic speech recognition of The Holy Qur'ān recitations," *Int. J. Speech Technol.*, vol. 24, no. 4, pp. 1017–1032, Dec. 2021, doi: <https://doi.org/10.1007/s10772-021-09853-9>
- [4] A. Ali, S. Naeem, S. Anam, and M. M. Ahmed, "Current state of artificial intelligence (AI) in oncology: A review," *Curr. Trends OMICS*, vol. 3, no. 1, pp. 1–17, June 2023, doi: <https://doi.org/10.32350/cto.31.01>
- [5] M. Zubair, "Machine learning based biomedical image analysis and feature extraction methods," *J. Appl. Emerg. Sci.*, vol. 13, no. 1, pp. 31–39, June 2023, doi: <http://dx.doi.org/10.36785/jaes.131551>
- [6] A. Ali *et al.*, "Connecting Arabs: bridging the gap in dialectal speech recognition," *Commun. ACM*, vol. 64, no. 4, pp. 124–129, Apr. 2021, doi: <https://doi.org/10.1145/3451150>
- [7] J. Younes, E. Souissi, H. Achour, and A. Ferchichi, "Language resources for Maghrebi Arabic dialects' NLP: A survey," *Lang. Resour. Eval.*, vol. 54, no. 4, pp. 1079–1142, Dec. 2020, doi: <https://doi.org/10.1007/s10579-020-09490-9>
- [8] A. O. Salau, T. D. Olowoyo, and S. O. Akinola, "Accent classification of the three major nigerian indigenous languages using 1D CNN LSTM network model," in *Advances in Computational Intelligence Techniques*, S. Jain, M. Sood, and S. Paul, Eds., Singapore: Springer, 2020, pp. 1–16, doi: [https://doi.org/10.1007/978-981-15-2620-6\\_1](https://doi.org/10.1007/978-981-15-2620-6_1)
- [9] I. Albayrak, "Revisiting the meaning of the divine preservation of the Qur'an: With special references to verse 15: 9," *Religions*, vol. 13, no. 11, Art. no.1064, Nov. 2022, doi: <https://doi.org/10.3390/rel13111064>

- [10] G. Krishna, C. Tran, M. Carnahan, Y. Han, and A. H. Tewfik, "Generating EEG features from acoustic features," in *28th Eur. Signal Process. Conf.*, Jan. 2021, pp. 1100–1104, doi: <https://doi.org/10.23919/Eusipco4796.8.2020.9287498>
- [11] A. Ali, S. Anam, and M. M. Ahmed, "Shannon entropy in artificial intelligence and its applications based on information theory," *J. Appl. Emerg. Sci.*, vol. 13, no. 1, pp. 9–17, June 2023, doi: <http://dx.doi.org/10.36785/jaes.131549>
- [12] S. Naeem, "Network security and cryptography challenges and trends on recent technologies," *J. Appl. Emerg. Sci.*, vol. 13, no. 1, pp. 1–8, June 2023, doi: <http://dx.doi.org/10.36785/jaes.131546>
- [13] Y. Wang *et al.*, "Transformer-based acoustic modeling for hybrid speech recognition," in *ICASSP 2020-2020 IEEE Int. Conf. Acoustics, Speech Signal Proc.*, May 2020, pp. 6874–6878, doi: <https://doi.org/10.1109/ICASSP40776.2020.9054345>
- [14] S. Naeem and A. Ali, "Bees algorithm based solution of non-convex dynamic power dispatch issues in thermal units," *J. Appl. Emerg. Sci.*, vol. 12, no. 1, June 2022, doi: <http://dx.doi.org/10.36785/jaes.121540>
- [15] A. Ali and S. Naeem, "The controller parameter optimization for nonlinear systems using particle swarm optimization and genetic algorithm," *J. Appl. Emerg. Sci.*, vol. 12, no. 1, June 2022, doi: <http://dx.doi.org/10.36785/jaes.121538>
- [16] A. Ali, S. Naeem, S. Anam, and M. M. Ahmed, "Machine learning methods of IoT security and future application," *Proc. Pak. Acad. Sci. Phys. Comput. Sci.*, vol. 59, no. 3, pp. 1–16, Aug. 2022, doi: [https://doi.org/10.53560/PPASA\(59-3\)782](https://doi.org/10.53560/PPASA(59-3)782)
- [17] N. J. Ibrahim, Z. M. Yusoff, and Z. Razak, "Improve design for automated Tajweed checking rules engine of Quranic verse recitation: a review," *Int. J. Quranic Res.*, vol. 1, no. 1, pp. 39–50, 2011.
- [18] M. S. Abdo, A. H. Kandil, A. M. El-Bialy, and S. A. Fawzy, "Automatic detection for some common pronunciation mistakes applied to chosen Quran sounds," in *5th Cairo Int. Biomed. Eng. Conf.*, IEEE, Dec. 2010, pp. 219–222, doi: <https://doi.org/10.1109/CIBEC.2010.5716073>
- [19] S. Naeem, A. Ali, S. Anam, and M. M. Ahmed, "An unsupervised machine learning algorithms: Comprehensive review," *Int. J. Comput. Digit. Syst.*, vol. 13, no. 1, Jul. 2023, doi: <http://dx.doi.org/10.12785/ijcds/130172>
- [20] A. Ali, S. Naeem, S. Anam, and M. M. Ahmed, "A state of art survey for big data processing and NoSQL database architecture," *Int. J. Comput. Digit. Syst.*, vol. 4, no. 1, July 2023, doi: <http://dx.doi.org/10.12785/ijcds/140124>
- [21] P. S. P. Kumar, G. T. Yadava, and H. S. Jayanna, "Continuous Kannada speech recognition system under degraded condition," *Circuits Syst. Sig. Process.*, vol. 39, no. 1, pp. 391–419, Jan. 2020, doi: <https://doi.org/10.1007/s00034-019-01189-9>
- [22] S. A. Bukhari, A. Ali, and S. Naeem, "Promoting the leadership effectiveness in HEC organizations of southern Punjab, Pakistan," *UMT*

- Educ. Rev.*, vol. 5, no. 2, pp. 102–119, Dec. 2022, doi: <https://doi.org/10.32350/uer.52.06>
- [23] A. Ali, S. Naeem, S. Anam, and M. M. Ahmed, “Entropy in information theory from many perspectives and various mathematical models,” *J. Appl. Emerg. Sci.*, vol. 12, no. 2, pp. 156–165, Dec. 2022, doi: <http://dx.doi.org/10.36785/jaes.122548>
- [24] A. Ali, S. Naeem, S. Anam, and M. M. Ahmed, “Machine learning for intrusion detection in cyber security: Applications, challenges, and recommendations,” *UMT Artif. Intell. Rev.*, vol. 2, no. 2, pp. 41–64, Dec. 2022, doi: <https://doi.org/10.32350/icr.0202.03>
- [25] K. C. Yoong and K. S. Hong, “Malaysian English large vocabulary continuous speech recognizer: an improvement using acoustic model adaptation,” *IEM J.*, pp. 53–61.
- [26] K.-F. Lee and H.-W. Hon, “Speaker-independent phone recognition using hidden Markov models,” *IEEE Trans. Acoust. Speech Signal Process.*, vol. 37, no. 11, pp. 1641–1648, Nov. 1989, doi: <https://doi.org/10.1109/29.46546>
- [27] H. Prakoso, R. Ferdiana, and R. Hartanto, “Indonesian automatic speech recognition system using CMU Sphinx toolkit and limited dataset,” in *Int. Symp. Elect. Smart. Devic.*, IEEE, 2016, pp. 283–286, doi: <https://doi.org/10.1109/ISESD.2016.7886734>
- [28] S. K. Saksamudre, P. P. Shrishrimal, and R. R. Deshmukh, “A review on different approaches for speech recognition system,” *Int. J. Comput. Appl.*, vol. 115, no. 22, pp. 23–28, Apr. 2015,
- [29] G. Luo, P. Yang, M. Chen, and P. Li, “HCI on the table: Robust gesture recognition using acoustic sensing in your hand,” *IEEE Access*, vol. 8, pp. 31481–31498, Feb. 2020, doi: <https://doi.org/10.1109/ACCESS.2020.2973305>
- [30] M. Zubair, “A brief history of information theory by Claude Shannon in data communication,” *J. Appl. Emerg. Sci.*, vol. 13, no. 1, pp. 23–30, June 2023, doi: <http://dx.doi.org/10.36785/jaes.131550>
- [31] A. A. Kabir, “Memorizing the sacred in the digital age: Exploring Qur’an memorization experiences using physical & digital formats,” Master thesis, Univ. Maryland, College Park, Maryland, USA, 2021.
- [32] M. M. T. Nur, S. S. Dola, A. K. Banik, T. Akhter, and N. Hossain, “Voice recognition using machine learning and central database to enhance security system,” Graduation thesis, Depart. Comput. Sci. Eng., Brac Univ., Dhaka, Bangladesh, 2022.
- [33] A. Ali, B. N. Hashmi, A. Batool, S. Naeem, S. Anam, and M. M. Ahmed, “Machine learning based implementation of home automation using smart mirror,” *UMT Artif. Intell. Rev.*, vol. 2, no. 1, Art. no. 1, Mar. 2022, doi: <https://doi.org/10.32350/Umtair.21.002>
- [34] H. Strik, K. Truong, F. de Wet, and C. Cucchiari, “Comparing different approaches for automatic pronunciation error detection,” *Speech Commun.*, vol. 51, no. 10, pp. 845–852, Oct. 2009, doi: <https://doi.org/10.1016/j.specom.2009.05.007>

- [35] A. Al Hindi, M. Alsulaiman, G. Muhammad, and S Al-Kahtani, "Automatic pronunciation error detection of nonnative Arabic Speech," in *11th Int. Conf. Comput. Syst. Appl.*, 2023, pp. 190–197, doi: <https://doi.org/10.1109/AICCSA.2014.7073198>
- [36] S. M. Witt and S. J. Young, "Phone-level pronunciation scoring and assessment for interactive language learning," *Speech Commun.*, vol. 30, no. 2, pp. 95–108, Feb. 2000, doi: [https://doi.org/10.1016/S0167-6393\(99\)00044-8](https://doi.org/10.1016/S0167-6393(99)00044-8)
- [37] C. Cucchiariini, F. D. Wet, H. Strik, and L. Boves, "Assessment of dutch pronunciation by means of automatic speech recognition technology," in *5th Int. Conf. Spoken Lang. Proc.*, 1998, Art. no. 0751, doi: <https://doi.org/10.21437/ICSLP.1998-720>
- [38] M. Maqsood, H. A. Habib, S. M. Anwar, M. A. Ghazanfar, and T. Nawaz, "A comparative study of classifier based mispronunciation detection system for confusing," *Nucleus*, vol. 54, no. 2, pp. 114–120, June 2017.
- [39] M. Maqsood, "Selection of discriminative features for Arabic phoneme's mispronunciation detection," *Pak. J. Sci.*, vol. 67, no. 4, pp. 405–412, Dec. 2015, doi: <https://doi.org/10.57041/pjs.v67i4.607>