Analysis of Momentous Fragmentary Formants in Talaqi-like Neoteric Assessment of Quran Recitation using MFCC Miniature Features of Quranic Syllables

Mohamad Zulkefli Adam, Noraimi Shafie, Hafiza Abas, Azizul Azizan
Razak Faculty of Technology and Informatics
UniversitiTeknologi Malaysia
Kuala Lumpur, Malaysia

***Abstract*—The use of technological speech recognition systems with a variety of approaches and techniques has grown rapidly in a variety of human-machine interaction applications. Further to this, a computerized assessment system to identify errors in reading the Qur'an can be developed to practice the advantages of technology that exist today. Based on Quranic syllable utterances, which contain Tajweed rules that generally consist of Makhraj (articulation process), Sifaat (letter features or pronunciation) and Harakat (pronunciation extension), this paper attempts to present the technological capabilities in realizing Quranic recitation assessment. The transformation of the digital signal of the Quranic voice with the identification of reading errors (based on the Law of Tajweed) is the main focus of this paper. This involves many stages in the process related to the representation of Quranic syllable-based Recitation Speech Signal (QRSS), feature extraction, non-phonetic transcription Quranic Recitation Acoustic Model (QRAM), and threshold classification processes. MFCC-Formants are used in a miniature state that are hybridized with three bands in representing QRSS combined vowels and consonants. A human-guided threshold classification approach is used to assess recitation based on Quranic syllables and threshold classification performance for the low, medium, and high band groups with performances of 87.27%, 86.86%and 86.33%, respectively.**

Keywords*—Speech processing; MFCC-Formant; Quranic recitation assessment; human-guided threshold classification

I. INTRODUCTION

The recitation of the Qur'an, which although uses Arabic words, however, is quite different from the recitation of ordinary Arabic texts. This is due to the presence of certain pronunciation rules (Tajweed) which must be followed during the recitation [1]. As a result, it can be agreed that those who are Arabs and practice Arabic are also required to learn pronunciation that conforms to Tajweed rules while reading the Quran. Tajweed rule basically emphasizes the correct and accurate pronunciation, which is called Makhraj (or plural Makhaarij). It involves the articulation point of each letter and together with determining the specific quality or characteristics (Sifaat) of each letter that distinguishes it from other sounds.Most of the applications provide Al Quran contents in text, audio and video formats without interactive tools to perform assessment of recitation.

Speech processing is widely used in human-computer interaction. Speech signals rich in speech information can be utilized by frequency modulation, amplitude modulation and time modulation that carry various components such as resonance movement, harmonic, tone intonation, force and even time. The research may lead to the approach of spectral energy and its temporal structure that can be used in speech processing in the recitation of the Quran. The measurement of the parameters and features extracted will be used to capture the nature of the speech and the parametric features to reveal the errors of Quranic recitation based on Tajweed measured from the likelihood of the parametric features.

The speech signal properties are used to be the main reference in the Quran recitation assessment computing machine, where the same method was developed and demonstrated in the Intelligent Quran Recitation Assistance (IQRA) computational engine proposed in the study presented in this paper. Firstly, the unique and salient features are identified, investigated and used to represent the digitized Tajweed rules that embedded in the recited syllable of particular Quranic word. This is then creatively and experimentally led to the creation of extractor and classifier design to underpin the task of dissimilarity grouping of Tajweed rules, where the assessment will take place. The main concern of this paper is to reveal the analysis process of the significance (momentous) level of the miniature features (fragmented formants) in producing the digital representation of the Tajweed rules (based on the syllable). By strategically using the threshold approach in the experiments, the conventional Talaqi-like approach seemingly realized digitally and formed the new modern (or neoteric) assessment.

In the remaining of the sections, the content of the paper is divided into various sections for the purpose of conveying the understanding of the proposed problem and solution. Section II discusses the signification of fragmented formants (each of several frequency bands) or momentous fragmentary formants that derived from the Mel Frequency Ceptral Coefficients (MFCC). This is then followed by Sections III and IV for the experiment and human-guided results, respectively. The outcome of the paper is concluded in Section V with comments and recommendation.

II. Momentous Fragmentary MFCC-FormantS

Although speech recognition techniques have evolved drastically and have begun to improve in application construction, they are still the most challenging method to analyse spoken language based on pattern recognition or machine ability in learning pattern development more interactively. Speech recognition commonly used in Arabic and Quran recitation are such as Arabic coding and synthesis research, dialect detection, speaker recognition, memorization and sentence retrieval. A large number of analyses use word or sentence utterance approach techniques [2] to identify and evaluate from signal speech representation. Spoken or readings are present as a form of language because speech also contains basic acoustic sounds or also known as phonemes. Each phoneme sound released is usually influenced by a neighbourhood phoneme delivered with a syllable or word. The recitation of the Quran conveys words spoken with a certain rhythm that can be formulated as an acoustic-phonetic symbol and prosody. Challenges of developing such a system are centralized to the modelling of features extraction and matching process that significantly are able to describe the recitation errors and intelligently propose the Tajweed error detection. Acoustic phonetics symbols of Arabic language can be formed as consonants and vowels. Each of combination Arabic phonetic symbols can be represented as a syllable and word. There are six pattern combinations of vowels which are CV, CV: CVC, CVCC, CV:C, CV:CC, where C represents the consonant and V as a vowel[3] while V: as a long vowel. The approach of analysis concerns on sequences of voiced or unvoiced sound because of recitation are related to phonetic and prosody. The sequences are segmented in a series of frames and represented by formant frequencies[3]and [4]. The production of voice or speech involves the movement of air from the lung to vocal tract towards the lips. The combination of voice production mechanism produces a variety of vibration and spectral-temporal composition that produce different speech sound. Apparently, the Arabic phonetic sound was produced from the specific articulation places or regions in the vocal tract. Speech or voice response is produced through the vocal tract filter that is characterised by series of formant or resonant frequencies [5].

The sound spectrum can be represented by formant frequencies which show the greater intensity of sound quality. The quality of sound is greatly shown using formant frequencies, especially the characteristic of sound of the consonant [6]. In this case, the formant frequencies of f_1, f_2, f_3 and $f_4$as illustrated in Fig. 1 are used as features to model the Quranic alphabet pronunciation [6]. Theoretically, the combination of formant frequencies of f_1, f_2, f_3 and $f_4$from speech production should be able to describe the characteristics of the letters during pronunciation for each of the 28 hijaiyah letters (Arabic letters).Furthermore, the changes of formant frequencies of f_1, f_2, f_3 and f_4 can be used to represent the characteristic of vowel, consonants and its combination [7][8]. There are large number of techniques used in speech processing and feature extraction can be used such as Mel Frequency Cepstral Coefficient (MFCC), PLP and LPC techniques [9]. MFCC is the most popular feature extraction technique compared to other techniques because it is related to the human auditory system [10]. In addition, MFCC can produce better accuracy with less computational complexity.

The momentous fragmentary MFCC-formant frequency is experimentally invented and introduced for representing the syllable feature with detail analytical approach. This is done by dividing the MFCC and its derivative feature into Band-1, Band-2 and Band-3 for representing the formant frequency ranges of low frequency, medium frequency and high frequency respectively. Each band has been broken into four (4) co-efficient of MFCC which represent the frequencies from the filter triangular bank. Each of the co-efficient derives a sequence of frames of power energy. The value energy in every frame is basically the total energy from multiple filters in the MFCC. The combination of frames from the first frame to the 'n = 1, 2, 3, …' frame is the concatenated $MFCC_n$ and its derivatives ($\Delta MFCC_n$ and $\Delta\Delta MFCC_n$). The co-efficient and their appropriated power energy frames can be considered as the miniature feature that will characterize the respected syllable feature. This approach takes into account the evaluation of vowel and consonant features in syllable pronunciation based on low, medium and high formant frequency ranges. The selection of band is also closely related to vowel and consonant of phoneme speech spectrum. Table I shows how the three bands have been fragmented by dividing the selected range of frequencies based on experiments. Vowel and consonant that are categorized as voiced phoneme have high power energy characteristics. In Band1, the MFCC coefficient (C1, C2,…, C12) as described in Table I are categorized based on the formant frequency range. The formant of f_1 which has coefficients C1, C2, C3 and C4 is indicative of High-Voiced Vowel and Low-Voiced Consonant Characteristic. While band2 represents the formant frequency of f_2 and $f_3$are represented by the coefficient C5, C6, C7 and C8 indicate the characteristics of High-Voiced Consonant and Low-Voiced Vowel Characteristic. The final category is band3 for the formant range, $f_4$and above. This category is represented by C9, C10, C11, and C12 which exhibit Voiceless and Low-Voiced Consonant Characteristic features. For the experiments conducted, band1 and band2 are very practical to reveal information about vowels, while information about sound consonants is more appropriately revealed on a combination of band1, band2 and band3. Band3 is therefore used to reveal voiceless consonant information. Table I lists the frequency ranges for the 3 bands and its coefficients.

Each band shows the characteristic of formant frequency of syllable pronunciation that is produced from the vocal tract filter response[11].

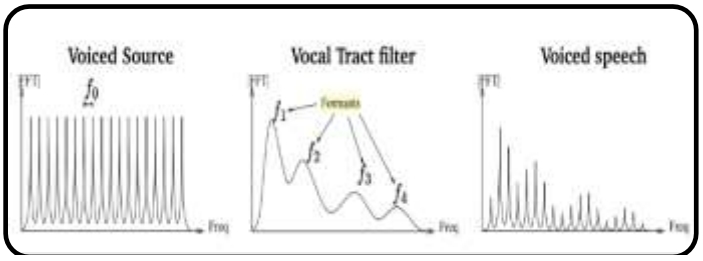

Fig. 1. The Source and Filter Response for a Typical Vowel Sound [5].

TABLE I. THE BAND OF FREQUENCY RANGE

Fragmentary Frequency	Band1 (300Hz-1077Hz)	Band2 (861.3Hz-2089Hz)	Band3 (1787Hz-3704Hz)
Syllable Pronunciation	High-Voiced Vowel and Low-Voiced Consonant	High-Voiced Consonant and Low-Voiced Vowel	Voiceless and Low-Voiced Consonant
MFCC-Formant co-efficient	C1, C2, C3 and C4	C5, C6, C7 and C8	C9, C10, C11 and C12

III. EXPERIMENT

The structure of IQRA implementation for the Al-Quran recitation assessment has been divided into three (3) stages as shown in Fig. 2. The first stage is data acquisition andpre-processing. The second stage is feature's extraction and last stage is Human-Guided Threshold classification.

The assessment algorithm of Quranic recitation uses tactical hybrid methodical DSP approaches by combination of various machine learning [12] conventional approaches. The flow of proposed computational engine for Quranic recitation assessment is described in the following sub-sections.

A. Data Acquisition and Pre-Processing

Data acquisition is recorded from numerous Malay reciters of various backgrounds. This includes male and female of Malay ethnic, ranging from the age of between 20 to 65 years old. There are two categories of selected reciters which are experts and learners. This Quranic Recitation Speech Signals (QRSS) originally also contained unwanted audio such as noise or any surrounding audio that is difficult to predict. However, the signal compensation method is used to eliminate these unwanted signals, which include such as the 60Hz Hum AC-DC signal [13], the silent signal [14], breaths sound signal, clicks and pops sound [9] that can interfere the performance of the computing engine.The wav format is a commonly recorded audio data format, for example with 16 bits, 44,100 samples [9] and uses mono channels.

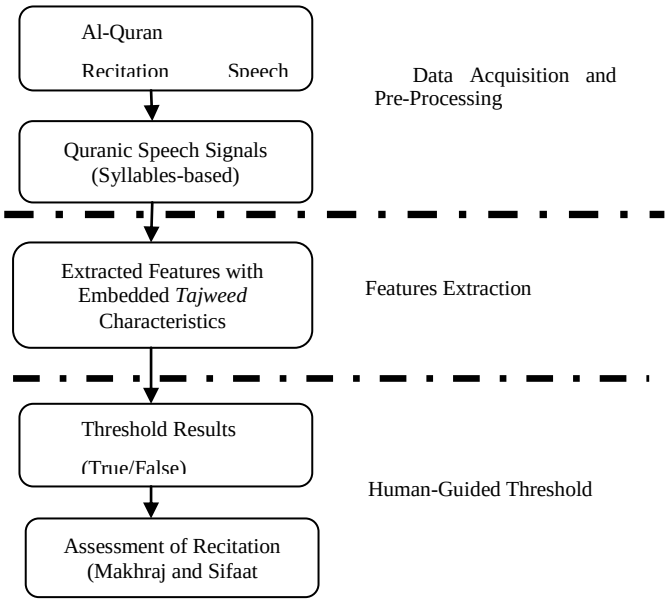

Fig. 2. The Structure of IQRA Implementation.

The main aim of signals initialization is to prepare the signals with several selected techniques that should be able to enhance signals representation. The steps of initialization are start-end point detection [14], pre-emphasis [15] and amplitude normalization. The end point detection is used to define the start point and end point of Quranic speech signals. Each of learners or experts have different start point and end point while do recitation. Combined zero crossing and short term energy function are used to determine start point and end point [16]. Therefore, the amplitude normalization is used to compensate the speaker health condition, age and gender and change the amplitude range between 0 and 1. Meanwhile, pre-emphasis converts the QRSS to the higher frequency with the co-efficient of 0.95. There will be more information can be extracted by converting the signal into high frequency spectrum as compared to the one in low frequency.

In confronting the variability and complexity of the continuous QRSS, the recitation of the experts and learners should be parameterized by a single warp factor. Based on vocal tract speech production, the air flow of speech production among reciters is differently delivered and it's involved of Vocal Tract Length (VTL). VTL is varied across different reciters around 18cm and 13cm for males and females, respectively. The positions of formant frequency are inversely proportional to VTL, and the formant frequency can vary around 25% [17]. The main purpose of speaker adaptation is to get the same rhythm, tone and length between expert and learner QRSS that can be compared in same word/utterance articulation from the Vocal Tract Length Normalization (VLTN)[18]. The DTW is used to warp QRSS energy of speech in the same length of recitation in the time series frame.

B. Feature Extraction and Prediction Model

The speech signal is basically a non-linear signal and needs to be handled with systematic processing. Thus, in this paper, the approach of Mel-Frequency Cepstral Coefficients (MFCC) and formant frequencies features (MFCC-Formant) are selected to reveal the characteristic of syllables by manipulating the power energy. The speech signals are segmented by time frame and also by frequency domain to derive the cepstral coefficients. The MFCC-Formant-like features are used as an acoustic model to indicate the pattern similarity and dissimilarity of Al-Quran recitation. The characteristic of the shape of the energy spectrum can be aligned as an acoustic model of Al-Quran recitation which represents the energy, rhythm and tone. The significant miniature feature for cepstrum energy and its derivative feature are extracted with the aid of cepstral analysis [16].

MFCCs have been widely used in the field of speech recognition and have successfully demonstrated dynamic features as they extract linear and non-linear properties of signals. MFCC and its derivatives (Δ and ΔΔ MFCC) are formed and grouped together to represent the transformed syllable. This QRSS is produced from the articulated speech production that consists of information of rhythm and intonation energy. ΔMFCC is known as delta coefficient (differential coefficient), while ΔΔMFCC is known as shift delta coefficient (acceleration coefficient) where both show the properties of trajectory of power energy between

segmented frame of syllable representation. MFCCs are more general acoustic features which are largely used in the systems that applied Good of Pronunciation (GoP) [19]. Firstly, use the Hamming window to find the magnitude of the signal by using the Fourier Transform. Secondly, map the power spectrum in frequency domain that obtained from the mel scale by using the triangular overlapping windows filters. Then, take logarithm power at each mel frequency and apply Discrete Cosine Transform (DCT) of mel log powers as if it were a signal. Lastly, MFCC represents the amplitude of the resulting spectrum. The block diagram of MFCC is show in Fig. 3.

Furthermore, Fig. 3 depicts the step and design approaches of feature extraction method for MFCC, ΔMFCC and ΔΔ MFCC then represent as MFCC-Formant miniatures features. In step 1, spectral analysis is used to determine the frequency formant content of the arbitrary signals of QRSS. The overlap frame that uses the hamming window is used to reduce the spectral leakage effect. On the side of hamming window, lobe is overlapped, and the main lobe captures the characteristic of spectral energy by using the Discrete Fourier Transform (DFT). The selection of hamming window is performed because of its least amount of distortion. The frame size must be controlled and not too large in order to prevent the QRSS syllable properties from being too much across the window, thus affecting the resolution of time, whereas if the frame size is too short, the resolution of the narrow-band component will be sacrificed, and this will adversely affect the frequency resolution. A large number of previous experiments using MFCC have stated that frame measurements for spectrograms preferably between 20ms and 40ms to optimize a sample sufficient to obtain reliable spectrum estimates and depend on the length of utterance. The frame size of 20ms and the frame shift of 10ms also have shown reliable spectrum estimation [21]. In this experiment, the chosen shape of spectrogram is framed between 25ms and frame shift is 10ms based on phoneme formant representation.

In step 2, the Mel scale is used to obtain the power spectrum for each frame. This can be done by using a triangular window filter where each of them is not the same size in terms of amplitude. The amplitude decreases with increasing frequency range, and this is to get the characteristics of low frequency and high frequency that can be heard by the human ear. The human ear is basically more sensitive to low frequencies.

In step 3, the logarithm power at each of filters is measured for every segmented frame. Thus, each of bin per frame per filter holds the log-energy for each filter channel. In the experiment done in this thesis, 20 numeric values are obtained for each frame at the output. The outputs are stored in a matrix form with the number of row represent the frame (size frame of QRSS syllable) and the number of columns equal to 20 (which is the number of filters in the filter bank).

In step 4, DCT converts the power spectrum log generated by the mel scale in the frequency domain to the time domain. The DCT will rearrange the co-efficient cepstral from small order to a large sequence based on the evaluation of cosine signal characteristics.

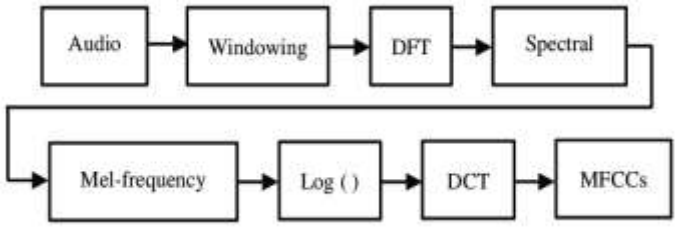

Fig. 3. The Block Diagram of MFCC [20].

In step 5, there are 13 MFCC-Formant coefficients generated from the QRSS syllable but only 12 coefficients are selected. The first coefficient (C0) representing the natural frequency (Pitch property) of the syllable indicates the amount of power energy but, not included in the analysis in this thesis. There are only from C1 to C12 MFCC-formant co-efficient are used in the analysis and taking the frequency of the band between 300Hz and 3700Hz. The 12 delta (ΔMFCC) and 12 ΔΔMFCC were concatenated together to represent the MFCC-Formant features of each QRSS syllable.

The speech signal is required as a stationary signal to estimate the parameters. The stationery signals were parameterised as features coefficient in such a manner before measuring the similarity through the matching or recognition process. Recitation of Al-Quran commonly can be assessed or evaluated using non-phonetic or phonetic transcription. In this paper, non-phonetic transcription is an approach by designing the prediction model without reference set of transcription. The parameter estimation algorithm of model prediction is estimate by using MLLR (Maximum Likelihood Linear Regression). These algorithms are integrated with GMM to classify the feature pattern as statistical model approach. These GMM statistical models have their characteristic which represent the signal characteristic as a static pattern [16]. The MLLR computes a set of transformations which reduces the mismatch between an initial model set and the adaptation data [22].

Parameter estimation is used to represent the acoustic model based on MFCC-formant-liked features and is designed to measure the similarity and dissimilarity (likelihood) of syllable pronunciation. The machine learning approach has great attention on parameter estimation in speech processing as data modelling. The QRAM is obtained by establishing a few tasks and methods to be applied. The fragmentary MFCC-formant features are proposed and modelled by using the Gaussian Mix Model (GMM). The GMM is a probabilistic model to represent the subpopulation and works well with the parameter estimation strategy. Generally, GMM is one of statistical-based clustering methods and an effective model that capable of achieving the high identification accuracy for short utterance length. Although MFCC is not robust to noise, the model-based approach used in this thesis able to eliminate the noise by the cancellation performed by Maximum Likelihood Estimation (MLE). MLE is a standard approach to estimate the model parameters from the sampling data. MLE is measured based on Expectation Maximization (EM) for parameters estimation approach. The EM algorithm is an iteration method to find the MLE of latent of hidden variables. The estimated parameters based on mean, covariance and

weight indicated the similarity and dissimilarity of each syllable pronunciation for every learner or reciter. The expected mean, covariant and weight of GMM for 4-Dimensional data are figured out by EM as mentioned before, where the EM algorithm is not possible to optimize the log likelihood of $\log[p(x|\lambda)$ directly with respect to λ. This means that the observation data, X= x1, x2,…,xD can be introduced by the discrete random variable, Z= z1, z2, …, zD and model parameters $\lambda=\{\overline{w}_i, \bar{\mu}_i, \bar{\Sigma}_i\}$. The log likelihood of model λ is given by

$$\sum_{d=1}^{D} \log p(x_d|\lambda) =$$

$$\sum_{d=1}^{D} \log p\left(x_d, z_{d,}|\lambda\right) - \sum_{d=1}^{D} \log p\left(z_d, x_{d,}|\lambda\right) \quad (1)$$

The expression also impresses that the statistical distribution of 4-dimensional observation of MFCC data can be clustered into 4-cluster of GMM. When the model of λ from different observation for different reciters is considered, each model will calculate the MLE parameters to represent the likelihood among the reciters for different band of MFCC. The parameters are trained as unsupervised classification. This model is designed by combining 4 GMM clusters using 4 dimensional fragmentary MFCCs to find the MLE that represents the data distribution for each of these frames. Furthermore, this model represents the sequence of MFCC-Formant sample frames that are considered parametric distribution models. The resulting MLE parameters show the maximum data calculated from the GMM model generated from the data that have been observed. This parameter is defined as a blueprint for the model. In avoiding the GMM overfitting, Bayesian Information Criterion (BIC) is used to estimate the reasonable amount of data prediction done by GMM. For instance, if the BIC value is much lower, the model is considered better in predicting data. BIC is an asymptotically optimal method for estimating the best model using only sample estimates [23]. BIC is defined as

$$BIC = -2 \ln I(x, M) + k \ln(n) \quad (2)$$

where x are the sample data, l(x,M) is the maximized likelihood function under a model M. While k is the number of estimated parameters, and n is the sample size.

The statistical clustering GMM Model approach is used to measure the similarity and dissimilarity of QRAM by estimating the maximum likelihood of fragmentary band of MFCC miniature features. It is a prototype-based algorithm which is consists of the feature vectors and representing as a mixture of Gaussian distribution. A mixture model is a probabilistic model for representing the presence of subpopulations within an overall population. The mixture shows probability distribution of parameters and can represent as the number of mixture components approaches to infinity. However, the appropriate number of mixtures must be determined for each model so that the mixtures are able to show the best distribution for the parameters or data where the distribution shows the characteristics of the parameters. Thus, the data will be segmented based on similarities or differences between observations in the dataset by 4-mixtures GMM as shown in Fig. 4. The similarities or differences are represented by the maximum likelihood estimation (MLE) as illustrated in Fig. 5. Each model of recitations is a set of model parameters with estimated mean vector, covariance matrix and mixture weight. Each of parameter models is trained as unsupervised classification by using the expectation maximization (EM).

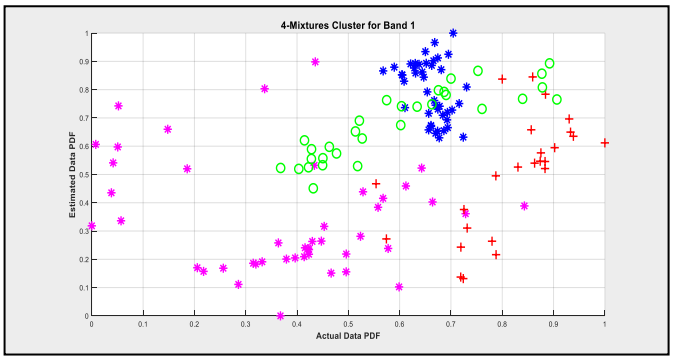

Fig. 4. The 4-Mixtures of GMM.

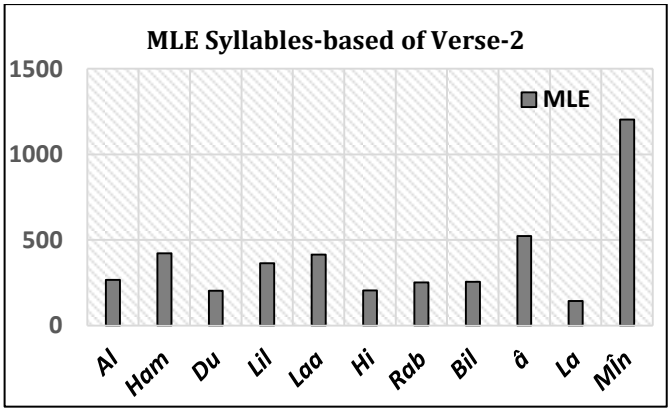

Fig. 5. The MLEs Value of Quranic Syllable-based Verse-2.

C. Human-Guided Threshold Classification

The computational engine score threshold process is used to evaluate the similarity and the dissimilarity based on human-guided threshold classification. This successful threshold process shows conventional Talaqi processes (experts evaluate the recitation by learners based on how to pronounce syllables in the verses of the Quran) are transformed to a machine evaluation approach. Computational engines must have salient features that can distinguish between correct and incorrect readings. Therefore, in determining the score, analysis of salient features, and matching process is used to obtain reading assessment based on the actual assessment by experts called a human-guided or Talaqi-Like assessment. Talaqi-like process has been used in the training phase and also the testing phase process in the computational engine. This is to ensure that the assessment by the expert is always included in the assessment made by the machine.

In this process, the MLE parameters are used as representations to each syllable recited by the learners. Initially, the MLE values from the expert readings were used as the initial reference in determining the initial threshold by assuming that all MLEs produced by the expert readings were within acceptable thresholds. After that the learners' reading is

assessed by the initial experts' threshold. Thus, the initial threshold will change to a new threshold after undergoing the training process by Human guided assessment (conducted by prominent expert). This process will be repeated until all MLE parameters have been evaluated by a prominent expert. Finally, the value of the threshold range has been completely obtained and can be used as a benchmark the reading made by the learners is correct or otherwise. The value of this Human-Guided Threshold classification will be tested in the testing phase and the performance is calculated.

IV. Result of Human-Guided Classification

A. Talaqi-Like Training Phase

In this training phase of the classification stage, the initial parameters of MLEs are taken from the calculation of 12 expert recitations. The starting point of training phase is when the input given to this designed system begins to create a change of pattern or minimum and maximum MLEs value that limits the correctness of a Tajweed in the reading of the Al-Fatehah chapter. This is seemingly caused by the changes of the acceptable lowest and highest values of that correspondingly due to the variability demonstrated by various reciters, but remains accepted (Acceptance Threshold) by the expert. The process of correcting (or training) the minimum and maximum values (threshold range) is firstly performed on the group of experts' MLEs data. This is the initial threshold range and used as reference values to be compared with the learner recitations. Secondly, the MLEs values obtained from the recited syllables of 40 learners are matched with the expert threshold range. Besides the setting of minimum and maximum values, the indication of True Acceptance (TA), False Rejection (FR) and False Acceptance (FA) of the calculated MLEs are counted and accumulated. Tabulates the MLEs values of syllables verse-2 of Al-Fatehah recited by 40 learners have been matched with the threshold range of expert's recitations. In the classification process performed in this experiment, the threshold range selected based on this expert indicates that most syllables are categorized as FR (False Rejection). This is logically agreeable and reasoned by the experts that most of the learners' performance has not been perfectly pronounced, but the Tajweed rules are acceptable.

The learning process for the machine evaluation to accurately perform is by allowing the human expert to guide the evaluation manually (Talaqi-Like approach). This is where the core of operations in the transfer of knowledge from human to a machine has taken place in the process of training the machine. Experts have individually altered the assessment performed by the machine in the case of (True) Rejection by the machine. This situation occurs by manually record or mark in the form that has been prepared for each learner. At the same time, the corresponding MLE values will be re-accepted as correct Tajweed recitation and assigned as True Acceptance (TA), which in turn will change the threshold range to new values (Minimum or Maximum). Fig. 6 shows the comparison of performance of classification for MLE band-1between initial expert threshold and Talaqi-Like threshold.

The acceptance of true recitation is based on three band threshold range categories of MLE, which indicate the similarity and dissimilarity for each syllable in Al-Fatehah verses, as compared with the adjusted reference threshold range. Similarity range indicates the acceptance and dissimilarity indicate unacceptance of recitation. The overall performance of acceptance recitation threshold is higher 80% for all MLEs. It shows that the MLEs parameter estimation can used indicator to assess the Quranic recitation assessment. However, the testing phase is used as validation stage to proof the experiment of Quranic assessment reliability. Fig. 7 shows the performance of Human-Guided Threshold classification based on true acceptance (TA) in training phase.

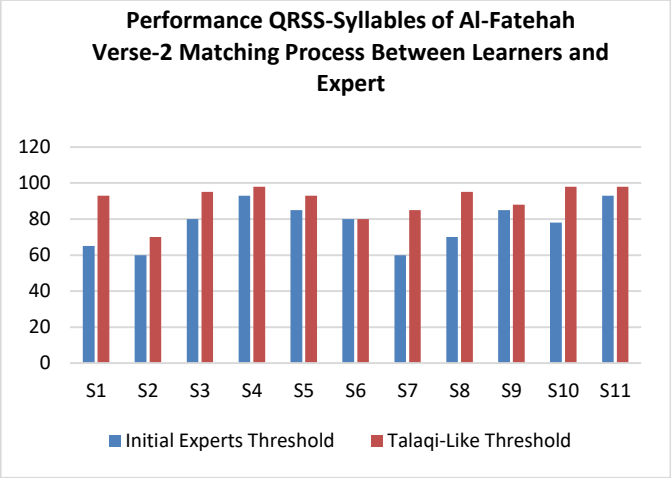

Fig. 6. Comparison Performance between Initial Experts and Talaqi-Like Threshold.

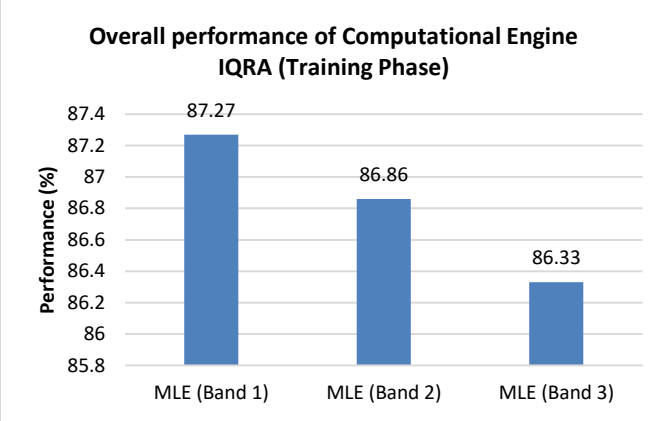

Fig. 7. Performance of IQRA in Training Phase.

B. Talaqi-Like Testing Phase

In the testing phase, the main objective is as linked from the training phase, which is to test the computational engine that has been designed in the context of reliability of the miniature salient feature, extractor and classifier. The trained range of MLEs is used to assess the performance of test data. Each syllable is tested according to the threshold determined based on MLE Band1, Band2 and Band3 (Human-Guided Threshold range). A total of 40 different learners from the training phase took their readings and the readings of each syllable in Al-Fatehah were extracted and matched with the reference MLEs from the training phase. Each test data is also evaluated manually by an expert and the performance of the reading truth that refers to Tajweed rules is calculated in a

technical context, namely, true / false positive acceptance (TP and FP), false rejection (FR) and false acceptance (FA). The comparison of errors will be made and analyzed between the machine evaluation and human evaluation. From here, the performance of the machine in terms of performing as an evaluator is then measured with respect to human expert performance.

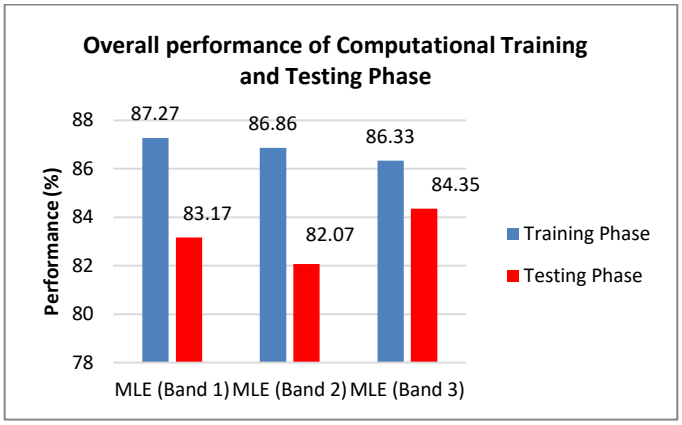

Fig. 8. IQRA Computational Training and Testing Phase Performance.

Based on Fig. 8, the classification performance using the threshold method for these MLEs parameters can be used to evaluate syllable-based Al-Quran recitation that Tajweed rules are embedded in the syllable. With this performance, the conclusion that can be expressed is that each band-1, band-2 and band-3 MLEs are able to show the characteristics of vowel and consonant combinations in each syllable based on fragmentary frequencies. These features have shown impressive performance of over 80% for representing Tajweed rules based on Makhraj, attributes and also derivatives rules.

V. Result and Discussion

The True Acceptance (TA) indicates both true positive and negative of syllables recitation based on location of MLEs. True positive shows that the learners have MLEs parameters are in the range of independent assessment threshold. It also shows that learners pronounced the syllables correctly. While true negative show that the learners pronounced the syllables incorrectly but MLEs parameters are out from threshold range. In addition, FR shows that the learners pronounced the syllable correctly but the location of MLE is located out from the threshold range. While FA indicates the pronunciation of syllables is incorrect, but the parameters MLE are in the range of independent assessment threshold.

Referring to Fig. 9, the total of FRR is revealed as 81.98%, while the total of FAR is 18.02%. This data represents all 40 learners involved in this testing phase. The plotted graph depicts the value of accuracy in evaluating readings depends on the ability to isolate either FA or FR. This ultimately leads to the sensitivity of the selected threshold change, where a reduction in FA will result in an increase to FR. In a simple interpretation, it is important for a learner to prioritize accuracy in reading in full compliance with Tajweed law. Therefore, the reduction in FA is considered better although it will lead to an increase in FR.

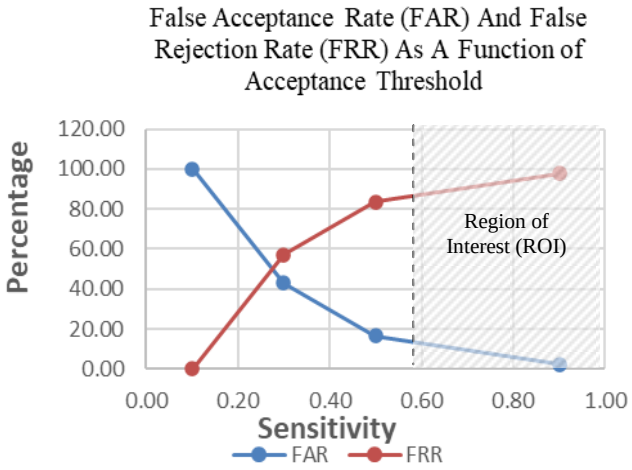

Fig. 9. FAR and FRR-The Function of Acceptance Threshold.

The goodness of pronunciation evaluation is used to evaluate the range and performance of recitation acceptance. The performance is measured by finding the threshold pattern for each syllable based on MLE parameters using GMM. In the training phase, there are two threshold processes involved, which are initial (or reference) threshold evaluation that based on expert recitation and expert-guided machine assessment. The overall performance of MLE Band-1, -2, -3 for Al-Fatehah verses are 86.33%, 86.86%, and 87.27%, respectively. The designed computational engine for IQRA (recitation assessment) system demonstrated through the use of fragmented frequency parameters (3 Bands) with the creation of salient miniature features of MLE along with machine learning has been implemented perfectly.

VI. Conclusion

Human-guided threshold classification process is studied and updated repeatedly based on observations given by prominent experts by looking for MLE parameters. This paves the way for the machine learning process through human-driven threshold values where the machine is able to assess learner recitation using MLE. The threshold is based on the probability or similarity of the MLE parameters of MFCC-Formant features for the syllables spoken by the reciters. The matching process practiced by this machine that uses human-guided threshold limit values can be interpreted as equivalent to the Talaqi approach as in the conventional evaluation process. In other words, the technological assessment in computer machines highlighted in this paper has successfully matched the way the assessment process of Quran recitation is done in conventional practice. Indeed, processes with highly systematic tactical and methodical approaches and techniques in combination with the role of human expertise and the advantages of the application of technology have been successfully demonstrated to produce a practical evaluation model.

Acknowledgment

Thank you to Universiti Teknologi Malaysia (UTM) for the financial support in sponsoring this research project, under

the program of UTM Enhancement Research Grant (UTMER).

REFERENCES

[1] Y. Mohamed, M. Hoque, T. H. S. Bin Ismail, M. H. Ibrahim, N. M. Saad, and N. N. M. Zaidi, "Relationship between Phonology, Phonetics, and Tajweed: A Literature Review," vol. 518, no. ICoSIHESS 2020, pp. 407–411, 2021, doi: 10.2991/assehr.k.210120.153.

[2] L. Marlina et al., "Makhraj recognition of Hijaiyah letter for children based on Mel-Frequency Cepstrum Coefficients (MFCC) and Support Vector Machines (SVM) method," 2018 Int. Conf. Inf. Commun. Technol. ICOIACT 2018, vol. 2018-Janua, pp. 935–940, 2018, doi: 10.1109/ICOIACT.2018.8350684.

[3] R. D. Kent and H. K. Vorperian, "Static measurements of vowel formant frequencies and bandwidths: A review," J. Commun. Disord., vol. 74, no. November 2017, pp. 74–97, 2018, doi: 10.1016/j.jcomdis.2018.05.004.

[4] S. Khairuddin et al., "Classification of the Correct Quranic Letters Pronunciation of Male and Female Reciters," in IOP Conference Series: Materials Science and Engineering, 2017, vol. 260, no. 1, doi: 10.1088/1757-899X/260/1/012004.

[5] M. N. Stuttle, "A Gaussian Mixture Model Spectral Representation for Speech Recognition," no. July, p. 163, 2003.

[6] S. Ahmad, S. N. S. Badruddin, N. N. W. N. Hashim, A. H. Embong, T. M. K. Altalmas, and S. S. Hasan, "The Modeling of the Quranic Alphabets' Correct Pronunciation for Adults and Children Experts," 2nd Int. Conf. Comput. Appl. Inf. Secur. ICCAIS 2019, pp. 1–6, 2019, doi: 10.1109/CAIS.2019.8769590.

[7] M. Farchi, K. Tahiry, S. Mounir, B. Mounir, and A. Mouhsen, "Energy distribution in formant bands for arabic vowels," Int. J. Electr. Comput. Eng., vol. 9, no. 2, p. 1163, 2019, doi: 10.11591/ijece.v9i2.pp1163-1167.

[8] K. Tahiry, B. Mounir, I. Mounir, L. Elmazouzi, and A. Farchi, "Arabic stop consonants characterisation and classification using the normalized energy in frequency bands," Int. J. Speech Technol., vol. 20, no. 4, pp. 869–880, 2017, doi: 10.1007/s10772-017-9454-9.

[9] T. Roy, T. Marwala, and S. Chakraverty, "Precise detection of speech endpoints dynamically: A wavelet convolution based approach," Commun. Nonlinear Sci. Numer. Simul., vol. 67, pp. 162–175, 2019, doi: 10.1016/j.cnsns.2018.07.008.

[10] M. Bezoui, A. Elmoutaouakkil, and A. Beni-Hssane, "Feature extraction of some Quranic recitation using Mel-Frequency Cepstral Coeficients (MFCC)," Int. Conf. Multimed. Comput. Syst. -Proceedings, vol. 0, pp. 127–131, 2017, doi: 10.1109/ICMCS.2016.7905619.

[11] N. Shafie, M. Z. Adam, H. Abas, A. Azizan, and K. Lumpur, "Sequential Classification for Articulation and Co-Articulation Classes of al-Quran Syllables Pronunciations Based on GMM- MLLR," AIP Conf. Proc., 2020.

[12] G. Aggarwal and L. Singh, "Comparisons of Speech Parameterisation Techniques for Classification of Intellectual Disability Using Machine Learning," Int. J. Cogn. Informatics Nat. Intell., vol. 14, no. 2, pp. 16–34, Feb. 2020, doi: 10.4018/ijcini.2020040102.

[13] L. R. Rabiner and R. W. Schafer, "MATLAB Exercises in Support of Teaching Digital Speech Processing," IEEE Int. Conf. Acoust. Speech Signal Process, pp. 2480–2483, 2014.

[14] M. Asadullah and S. Nisar, "A Silence Removal and Endpoint Detection Approach for Speech Processing 3 rd International Multidisciplinary Research and Information Technology," 3rd Int. Multidiscip. Res. Conf. 2016, no. September 2016, pp. 119–125, 2016.

[15] K. Livescu, P. Jyothi, and E. Fosler-Lussier, "Articulatory feature-based pronunciation modeling," Comput. Speech Lang., vol. 36, pp. 212–232, 2016, doi: 10.1016/j.csl.2015.07.003.

[16] N. Shafie, M. Z. Adam, S. Mohd Daud, and H. Abas, "A Model of Correction Mapping for Al-Quran Recitation Performance Evaluation Engine," Int. J. Adv. Trends Comput. Sci. Eng., vol. 8, pp. 208–213, 2019.

[17] L. Saheer, J. Dines, and P. N. Garner, "Vocal tract length normalization for statistical parametric speech synthesis," IEEE Trans. Audio, Speech Lang. Process., vol. 20, no. 7, pp. 2134–2148, 2012, doi: 10.1109/TASL.2012.2198058.

[18] N. Shafie, M. Z. Adam, and H. Abas, "Al-Quran Recitation Speech Signals Time Series Segmentation for Speaker Adaptation using Dynamic Time Warping," J. Fundam. Appl. Sci., vol. 10, pp. 126–137, 2018, doi: 10.4314/jfas.v10i2s.11.

[19] M. Maqsood, A. Habib, and T. Nawaz, "An efficient mispronunciation detection system using discriminative acoustic phonetic features for Arabic consonants," Int. Arab J. Inf. Technol., vol. 16, no. 2, pp. 242–250, 2019.

[20] H. C. Junho Son, Chon-Min Kyung, "Practical Inter-Floor Noise Sensing System with Localization and Classification," MDPI, no. August, 2019.

[21] M. Al-Ayyoub, N. A. Damer, and I. Hmeidi, "Using deep learning for automatically determining correct application of basic quranic recitation rules," Int. Arab J. Inf. Technol., vol. 15, no. 3A Special Issue, pp. 620–625, 2018.

[22] D. P. Lestari and A. Irfani, "Acoustic and language models adaptation for Indonesian spontaneous speech recognition," ICAICTA 2015 - 2015 Int. Conf. Adv. Informatics Concepts, Theory Appl., pp. 1–5, 2015, doi: 10.1109/ICAICTA.2015.7335375.

[23] J. Ding, V. Tarokh, and Y. Yang, "Model Selection Techniques: An Overview," IEEE Signal Process. Mag., vol. 35, no. 6, pp. 16–34, 2018, doi: 10.1109/MSP.2018.2867638.