Makhraj Recognition for Al-Quran Recitation using MFCC

Article *in* International Journal of Intelligent Information Processing · June 2013

DOI: 10.4156/ijiip.vol4.issue2.5

7 authors, including:

Nurul Wahidah Arshad
Universiti Malaysia Pahang
21 PUBLICATIONS **50** CITATIONS
SEE PROFILE

Lailatul Niza Muhammad
Universiti Malaysia Pahang
2 PUBLICATIONS **13** CITATIONS
SEE PROFILE

Hasan bin Ahmad
Universiti Malaysia Pahang
2 PUBLICATIONS **13** CITATIONS
SEE PROFILE

Rosnaaini Hamid
Universiti Utara Malaysia
12 PUBLICATIONS **29** CITATIONS
SEE PROFILE

Makhraj Recognition for Al-*Quran* Recitation using MFCC

[1]Nurul Wahidah Arshad, [1]Suriazalmi Mohd Sukri, [1]Lailatul Niza Muhammad, [2]Hasan Ahmad,
[1]Rosyati Hamid, [1]Faradila Naim, [1]Noor Zirwatul Ahlam Naharuddin
[1]Faculty of Electrical and Electronics Engineering
Universiti Malaysia Pahang
26600 Pekan, Pahang, Malaysia
wahidah@ump.edu.my, suriazalmi@gmail.com
[2]Centre for Modern Languages and Human Sciences
Universiti Malaysia Pahang
Lebuhraya Tun Razak, 26300 Gambang
Kuantan, Pahang, Malaysia

Abstract

This article presents a new application of recitation verification based on correct makhraj. Traditionally, people learn how to recite al-Quran correctly from an expert where it takes lots of time and effort. In this work, a new way to learn reciting the al-Quran is proposed in order to reduce the duration of learning from the expert. A system using Mel Frequency Cepstrum Coefficient (MFCC) as feature extraction and Mean Square Error (MSE) as a pattern matching technique is considered in order to develop a makhraj recognition system. An experiment has been setup to measure the system performance in terms of accuracy based on False Reject Rate (FRR) and Wrong Recognition (WR).

Keywords*: Speech Processing; Mel Feequency Cepstrum Coefficient (MFCC); Mean Square Error (MSE)*

1. Introduction

Al–Quran is the holy book of Muslims and the contents of the *Al-Quran* were written in Arabic. Most Muslims around the world knows how to recite *Al-Quran*, but not all Muslims can recite *Al-Quran* properly based on *makhraj* and *tajwid. Makhraj* and *tajwid* is an Arabic word for elocution. *Makhraj* is the correct position of the organs of speech in order to produce a letter so that it can be differentiated from others [1]. Meanwhile, t*ajwid* is the correctness of diction or proper pronunciation and technique during recitation [2]. It is compulsory for the Muslims to recite *Al-Quran* based on proper *makhraj* and *tajwid*. This is because, even a little difference of sound in an Arabic word can lead to a different meaning of the word. .

Basically, people will refer to the expert person in *makhraj* recitation. The expert will verify the recitation of *Al-Quran* and correct it if there are any mistakes in the recitation. Nowadays, there are several *Quran* learning software available in the market. Using these softwares, user only can listen and learn the correct pronunciation of the *Al-Quran*, but their pronunciation and recitation cannot be corrected like a teacher does because they themselves sometimes cannot detect their mistakes in recitation.

In this research, we develop a system using combination of the sound of *hijaiyah* letter as the input data to obtain the correct *makhraj* precisely by comparing the input to the sound of a single *hijaiyah* letter. *Hijaiyah* is an Arabic letter for *Al-Quran* Recitation. There is 29 basic *Hijaiyah* letter used in the Holy Quran. The main part on this paper is speech processing where we are using MFCC for feature extraction. It is the most critical step of speech recognition processing and directly affects the performance of the speech recognition system [3]. One of the most establishes and popular technique applied in speech recognition is MFCC for feature extraction of speech data [4]. The MFCC was chosen because of its robustness as discussed in [5]. It was derived from the Fast Fourier Transform (FFT) of the speech signal, where the frequency bands are positioned logarithmically on the Mel scale [6]. MFCC focuses on the human ear's non-linear frequency characteristic and the size of Mel frequency corresponds to the relation of actual frequency's logarithmic distribution on the whole and

accords with the human ear's characteristic [7]. The standard procedure for MFCC algorithm is as described below [8]:

i. Take short-time Fourier transform for the signal every 10 ms with 20 ms Hamming window,
ii. Map the power obtained in 1. into the mel scale using triangular overlapping windows,
iii. Compute the log-energy of each filter output,
iv. Applied Discrete Cosine Transform (DCT) to the filter bank output.

2. Related works

Recently, there are several research effort focuses on Arabic or Al-Quran recitation in terms of speech analysis and speech recognition such as in [1] and [9]. N. W. Arshad et. al focuses on noise removal in *makhraj* recognition using Normalized Least Mean Square (NLMS) Algorithm based on Adaptive Filter to search for the optimal solution. They only use 7 alphabets as samples, they are ا to ج. The speech processing method is used to obtain same waveform output from two different situations in MATLAB environment [1]. Successful utilization of RLS adaptive filter is proved by [10] with increased accuracy up to 98%. However, this filter requires high computational complexity and stability problem on tracking performance [11].

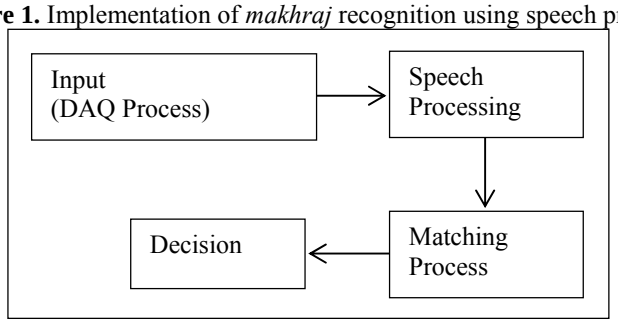
In [9] they propose a research to teach learners how to pronounce Arabic sounds and correct it if there are any mistakes. A speech database of the recited Quran is built where the sounds are labeled and segmented. The aim of this research is to create a set of labels that cover all the Arabic phonemes and their allophones. The set needs to include the sound system of the Classical Arabic (CA) and that of the Modem Standard Arabic (MSA) in addition to be flexible to include the sounds found in the Arabic dialects. The method for transcription has been applied to Quranic recitation to collect a sufficient Quranic speech database for training and testing. The database will be used to build the Computerized Teaching of the Holly Quran system in the Hidden Markov Toolkit (HTK) environment.

Yuan Yujin, Zhao Peihun and Zhou Qun have carried out a research about speaker recognition based on combination of Linear Prediction Cepstral Coefficient (LPCC) and MFCC [7]. The objective of this research is to prove that the combination of LPCC and MFCC as the feature extraction give the higher recognition rate. According to their paper, the key to speaker recognition is extracting speaker's personal audio traits. In this paper, LPCC and MFCC are used as the features extraction. Meanwhile, Vector Quantization (VQ) and DTW are used to recognize a speaker's identity in this system.

3. System implementation

Figure 1 shows the basic operation of *makhraj* recognition process; starting with data acquisition (DAQ) process, signal (speech) processing, feature extraction using MFCC and matching for decision purpose.

Figure 1. Implementation of *makhraj* recognition using speech processing



3.1. Data acquisition

The input speech was taken from people who are expert in m*akhraj* utterance between the ages of 21 and 23. The input speech was recorded in a sound proof room in order to minimize the disturbance of noise. Table 1 shows the several combination of two *hijaiyah* alphabet. Subjects will recite a combination for *hijaiyah* alphabet using mono microphone for 10 times and all the voices has been recorded using Audacity Version 1.3 Software in .wav format. The sampling frequency for each sample is 16 000 KHz.

Data taken are using recitation based on *Rasm 'Uthmani* narrated by Hafs bin Sulaiman for *Qira'at Asim*. All the data was validated by the certified expert during recording process.

Table 1. The combination of *hijaiyah* alphabet

Combination of *Hijaiyah* alphabet	**Pronounciation**
آب	ab
آت	at
آث	ath
آج	aj
آح	ah
آخ	akh
آد	ad
آذ	az
آر	ar
آز	azz

To develop a system that can verify the correct *makhra*j, the sample data is taken from the expert in recitation and these data are the reference data. These data are divided into two parts, one part is used for training phase and the remainder data will be used in testing phase.

3.2. Speech processing

The purpose of this stage is to make sure that only the desired signal will be processed. The step in speech processing are shown in Figure 2 which consist of silent detection, pre-emphasis, windowing, Fast Fourier Transform, mel frequency filter and DCT, where the MFCC algorithm is applied from the step 3 until the final step in Figure 2.

Silent detection is used to remove unvoiced signal. Meanwhile, the purpose of pre-emphasis is to emphasize higher frequencies. Equation 1 shows how the pre-emphasis work.

$$Y(n) = X(n) - aX(n-1) \quad (1)$$

Where $X(n)$ is the input of the signal, a is the percent of any one sample is presumed to originate from previous sample.

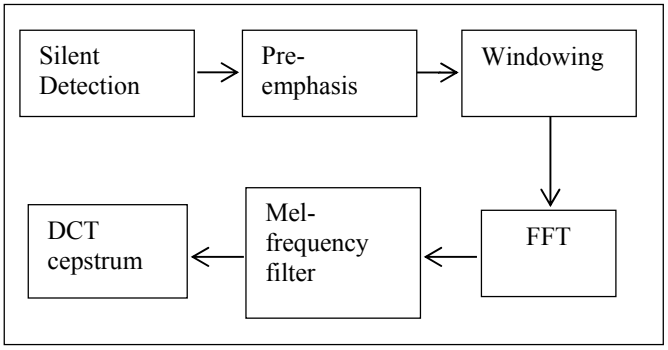


Figure 2. Detail process for speech processing stage

Windowing is used to make sure the continuity of the first and the last sampled speech in frames. There are several types of window that can be used for filtering such as Hamming, Hann and Blackman. The coefficients of windows are computed from the equation 2, 3 and 4.

Hamming Window :

$$w(n) = 0.54 - 0.46\cos(2\pi n/N) \quad , 0 \le n \le N \tag{2}$$

Hann Window :

$$w(n) = 0.5\,(1 - \cos(2\pi n/N)) \quad , 0 \le n \le N \tag{3}$$

Blackman Window :

$$w(n) = 0.42 - 0.5\cos(2\pi n/N) + 0.08\cos(4\pi n/N), 0 \le n \le N \tag{4}$$

Where n is number of sample and N represents the width, in samples.

Next, FFT is applied to obtain the magnitude frequency response of each frame and then the extraction process started. The purpose of feature extraction is to get the unique feature or coefficient for every input. The feature was extracted by using MFCC.

In order to get the coefficients of the input, the triangular window need to be determined. The reasons for using triangular bandpass filters because it smooth the magnitude spectrum such as the harmonics are flattened in order to obtain the envelope of the spectrum with harmonics. As a result, a speech recognition system will behave more or less the same when the input is same but different tones. Besides, the triangular filter also can reduce the size of features involved.

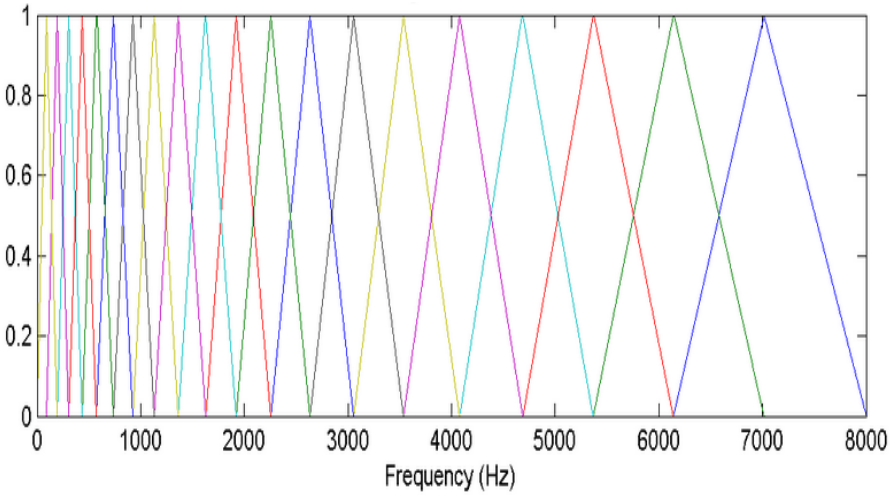


Figure 3. Triangular window

Figure 3 showed a set of triangular filters that are used to compute a weighted sum of filter spectral components so that the output of process approximates to a Mel scale. Each filter's magnitude frequency response is triangular in shape and equal to unity at the centre frequency and decrease linearly to zero at centre frequency of two adjacent filters. Then, each filter output is the sum of its filtered spectral components.

After the triangular filter bank generated, the mel frequency was computed by using equation 5.

$$f_{mel} = 2595 \log\left(1 + \frac{f}{700}\right) \tag{5}$$

Where f is the actual frequency in Hz.

The DCT was performed after calculate the mel frequency. Output from DCT consists of cepstral coefficient that represent feature for each input data.

3.3 Matching

We are using Mean Square Error (MSE) for pattern matching, where it measure the average of the squares of the errors. The MSE equation is shown in equation 6 where x is the coefficient value of a database input, while y is the coefficient value of a challenge data. Meanwhile, k is the number of filter such as 20. The purpose of MSE is to determine the mean square error between the coefficient value of the database and coefficient value of challenge data. From this MSE value, the threshold value is created in order to make sure the system verify the correct speech and reject false speech.

$$MSE = \sum (x - y)^2 / k \tag{6}$$

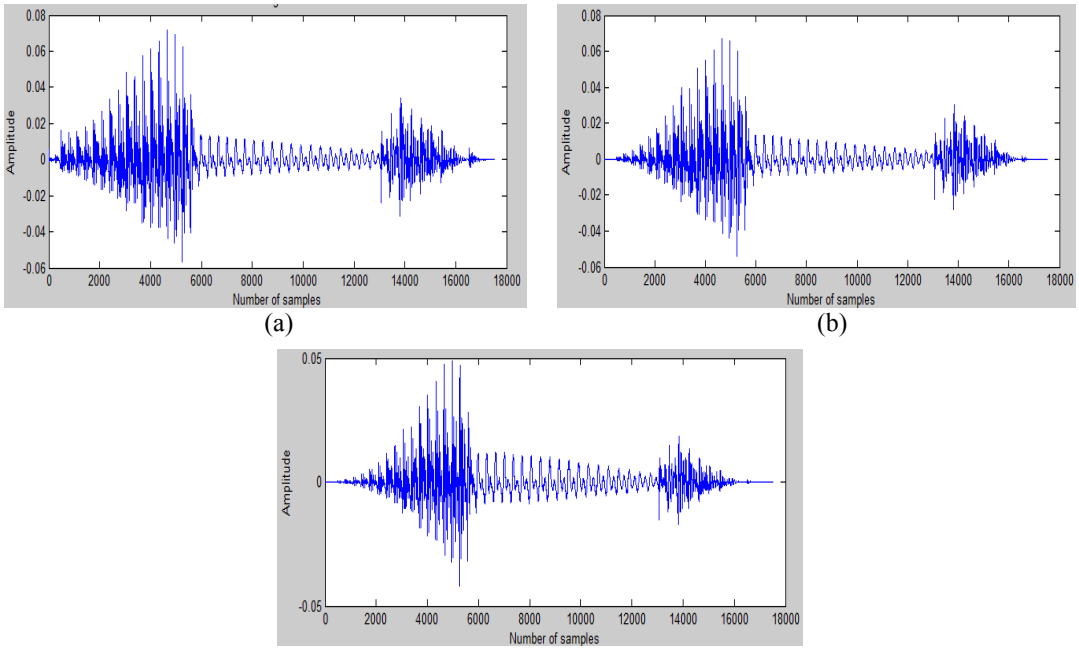
3.4. Experimental setup

In the experiment, ten experts in recitation including five males and five females were paid to use their voice in generating the data samples. Each person will repeat the same combination of *hijaiyah* alphabet for ten times, given a total number of data more than 2000 samples. About 20 data were use as a reference for training purpose. The remainder data were use as challenge data to to verify the validity and know the performance of the system.

This system is working in two different modes; one-to-one and one-to-many. In one-to-one mode, our database consist of one combined *hijaiyah* recitation but taken from different person, while in one-to-many, the database consist of all 26 combined *hijaiyah* recitation from different person.

4. Results and discussion

According to the algorithm of MFCC discussed earlier, Hamming window is applied to this algorithm. However, different windows are also used for reaching to the best performance of the algorithm. Figure 4 shows the windowed signal by using three types of window which are Hamming, Hann and Blackman window.

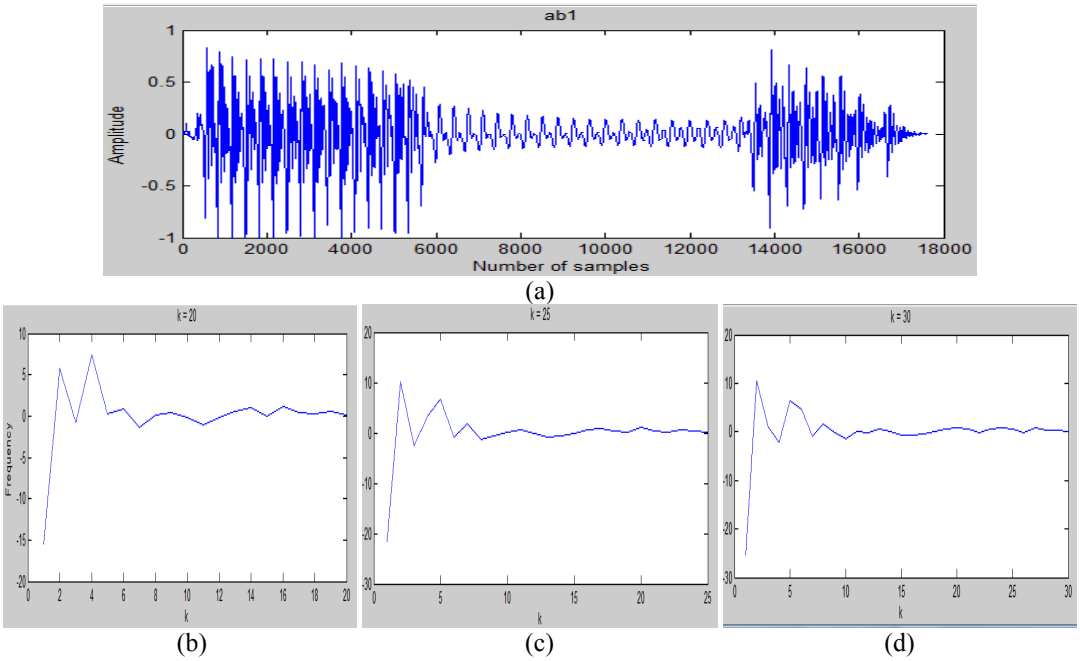


(a) (b)

(c)

Figure 4. Output signals of different window for 'ab' (آب) recitation by Person 1. (a) Hamming, (b) Hann and (c) Blackman

From Figure 4, we can see that for Hamming window gives the best output compared to Hann and Blackman window due to the retaining of the signal at the early number of samples. Using Hann and Blackman window, the output signal is affected due to the lost of information in the signal during the windowing process.



(a)

(b) (c) (d)

Figure 5. (a) The original signal, (b), (c), and (d) the significant coefficients for different value of filter for 'ab' (آب) recitation

To produce the cepstral coefficients, we take the first 20 low order coefficients. This is because they carry no significant information for the higher order coefficients which is if more than 20. Figure 5 shows the significant coefficient falls between 0 to 10, but we extend the value into 20.

The decision for this system is made after considering the minimum and maximum value of MSE. We use these minimum and maximum values from MSE as threshold value to recognize the correct *makhraj*. Figure 6 show one example how the threshold value was computed for 'ab' (آب) recitation.

FRR and WR were used to measure the system performance in terms of accuracy. FRR is the measurement of likelihood whether the system will incorrectly reject the correct *makhraj*. For example, if the input is 'ad' (آد) and the database consist of the same signals but taken from different persons, but the system verify it as incorrect recitation. FRR was used to measure the system performance for one-to-one mode. Meanwhile, the WR is the input that the system recognizes it as different recitation, where we calculate the WR for one-to-many system. For example, if the input is 'ab' (آب), but this system recognizes it as 'ath' (آت). Equation 7 and 8 show how we compute the system performance in terms of accuracy.

$$FRR = \frac{m}{n} \times 100\% \tag{7}$$

$$WR = \frac{p}{n} \times 100\% \tag{8}$$

Where m is the number of input that the system rejected, while p is the number of input that had been recognized different from the testing input and n is the total number of challenge input. While, Figure 7 shows the overall result from the first process until the final process.

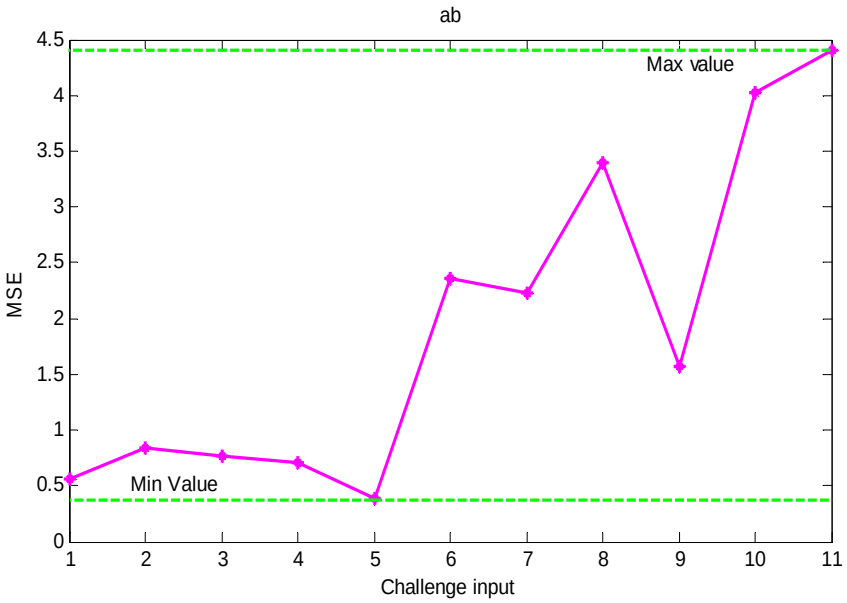


Figure 6. Threshold value for 'ab' (آب)

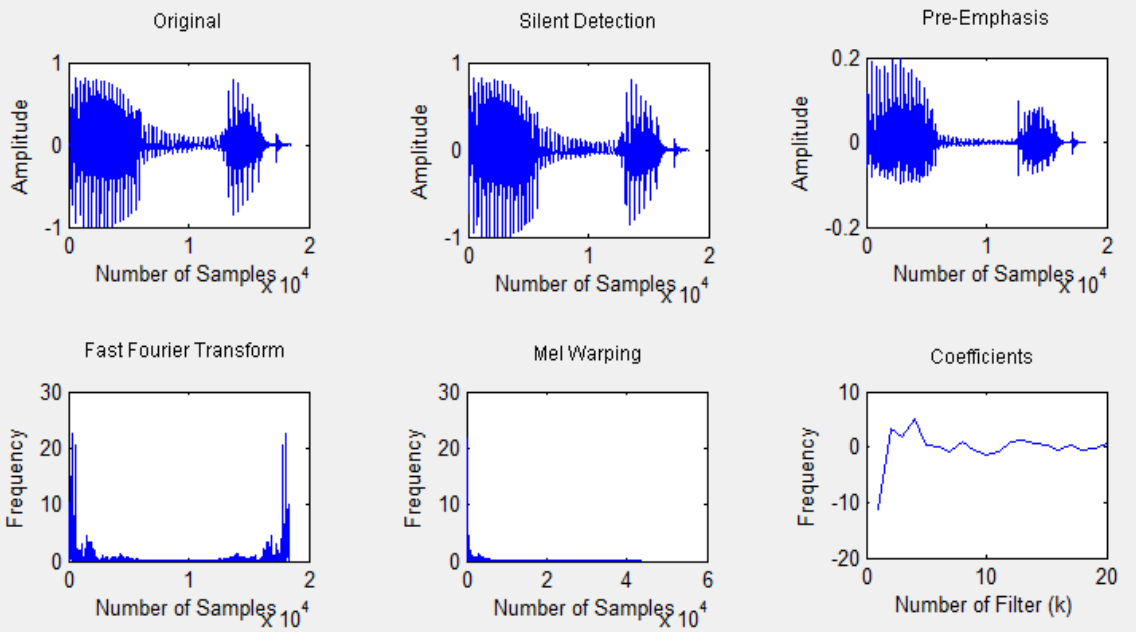


Figure 7. Result from overall process

Based on Table 2, the percentage of FRR for all recitation is 0% where it shows this system perform 100% accuracy which is high accuracy. While, the percentage of WR is very high for all the challenge input except for 'ab' (آب). This result happened because of the range of threshold value. For the certain signal, their range becomes a subset of another signal, so that the correct decision is hard to achieve.

Table 2. Percentage of System Performance

Challenge input		**One-to-One**	**One-to-Many**
		FRR (%)	*WR (%)*
ab	آب	0	0
ats	آت	0	90.9
ath	آث	0	100
aj	آج	0	100
ah	آح	0	100
akh	آخ	0	100
ad	آد	0	100
az	آذ	0	100
ar	آر	0	100
azz	آز	0	100

5. Conclusion and future work

A *makhraj* recognition system was build using speech processing technique. MFCC feature extraction is applied to this research to compute the significant coefficients for each input data. This system has successfully recognized the correct *makhraj* for one-to-one mode which give an average of accuracy of 100%. In the other hand, this system has an obstacle for one-to-many recognition as there is a need to extend process to do the matching process due to the simplicity of matching technique that was used.

There is still a room for improvement and some recommendation will be given for future work. More speech data should be taken like verses in the *al-Quran* so that this system can be implemented for *al-Quran* recitation. Besides, another technique for pattern matching should be tested for future work such as Dynamic Time Warping (DTW) and Hidden Markov Model (HMM).

6. Acknowledgement

This work is supported by University Research Grant RDU110367 and the data collection supported by Maahad Tahfiz Negeri Pahang (MTNP), Malaysia and Audio Lab of Centre for Modern Languages and Human Sciences, Universiti Malaysia Pahang.

7. References

[1] N.W.Arshad, S.N.Abdul Aziz, R. Hamid, R. Abdul Karim, F. Naim and N. F. Zakaria, "Speech Processing for Makhraj Recognition: The Design of Adaptive Filter for Noise Removal", In Proceeding of the IEEE Trans. on Electrical, Control and Computer Engineering (INECCE), pp. 323-327, 2011.

[2] Sonn, Tamara. "*Tajwid*". In Leaman, Oliver. *The Qur'an: an encyclopedia*. Great Britain: Routeledge, 2006.

[3] Huan Zhao, Yufeng Xiao, "A Novel Robust MFCC Extraction Method Using Sample-ISOMAP for Speech Recognition", International Journal of Digital Content Technology and its Applications (JDCTA), AICIT, vol. 6, no. 19, pp. 393-400, 2012.

[4] M. Hassan Shirali-Shahreza and Sajad Shirali-Shahreza, "Effect of MFCC Normalization on Vector Quantization Based Speaker Identification", In Proceeding of the IEEE Trans. on Signal Processing and Information Technology (ISSPIT), pp. 250-253, 2010.

[5] Rimah Amami, Dora Ben Ayed, Noureddine Ellouze, "An Empirical Comparison of SVM and Some Supervised Learning Algorithms for Vowel Recognition", International Journal of Intelligent Information Processing (IJIIP), AICIT, vol. 3, no. 1, pp. 63-70, 2012.

[6] N. Kamaruddin, A. Wahab and C. Quek, "Cultural dependency analysis for understanding speech emotion", International Journal of Expert Systems with Applications, ScienceDirect, vol. 39, no. 5, pp. 5115-5133, 2012.

[7] Yuan Yujin, Zhao Peihua and Zhou Qun, "Research of Speaker Recognition Based on Combination of LPCC and MFCC", In Proceeding of the IEEE Trans. on Intelligent Computing and Intelligent Systems (ICIS), pp.765-767, 2010.

[8] Nengheng Zheng, Tan Lee and P. C. Ching, "Integration of Complementary Acoustic Features for Speaker Recognition", International Journal of Signal Processing Letter, IEEE, vol. ASSP-14, no. 3, pp. 181-184, 2007.

[9] M.alghamdi, Y.O.Mohamed El Hadj and M.Alkanhal, "A Manual System to Segment and Transcribe Arabic Speech", In Proceeding of the IEEE Trans. on Signal Processing and Communications (ICSPC), pp.233-236, 2007.

[10] S.A.R. Al-Haddad, S.A. Samad, A. Hussain, K.A. Ishak and A.O.A. Noor, "Robust Speech Recognition Using Fusion Techniques and Adaptive Filtering," American Journal of Applied Sciences, Science Publications, vol. 6, no. 2, pp. 290-295, 2009.

[11] Georgi Iliev and Nikola Kasabov,**"**Adaptive Filtering with Averaging in Noise Cancellation for Voice and Speech Recognition," IEEE Trans. on Acoust., Speech, Signal Processing. Jan. 2002.