Iqra reading verification with mel frequency cepstrum coefficient and dynamic time warping

Adi Saputro Wibowo¹, I Dewa Made Bayu Atmaja Darmawan¹*
¹Informatics Department, Faculty of Mathematics and Natural Sciences, Universitas Udayana, Kampus Bukit Jimbaran, Badung, Indonesia

*dewabayu@unud.ac.id

Abstract. Al-Quran is a holy book of guidance from Allah SWT that was revealed through Prophet Muhammad SAW for Muslims. Muslims should learn to read Al-Quran properly and correctly. One of the methods in learning the Al-Quran is the Iqra method. Iqra method is learning in reading and writing the Al-Quran by using Iqra books delivered classically and individually. Learning methods will continue to be developed to adapt to current conditions. As technology develops today, the method of learning the Al-Quran has also developed, one of which is using sound. In this research, the writer will create a system that verifies the sound of reading the Iqra which is one of the basic methods in learning to read the Al-Quran. The system is designed using the Mel Frequency Cepstrum Coefficient (MFCC) method as a voice feature extension and the Dynamic Time Warping (DTW) method for matching the results of the voice features. In this study, the best accuracy in the Iqra reading sound was found to be 82% with 13 MFCC coefficients.

1. Introduction
Al-Quran is a holy book of guidance from Allah SWT that was revealed through Prophet Muhammad SAW for Muslims. Al-Quran was revealed not in the form of text, but the form of oral discourse. Al-Quran was revealed gradually over 30 years, so that it is easy to read, memorize, and understand its meaning. Educating and developing learning the Al-Quran is an obligation for Muslims. One of the methods of learning Al-Quran is the Iqra method. Iqra is learning in reading and writing Al-Quran by using Iqra books delivered classically and individually [1]. Learning methods will continue to be developed to adapt to current conditions. In the field of technology, many studies have been carried out to facilitate learning Al-Quran. One of the most frequent research in speech processing.

Based on this, the author will design a system that can facilitate Al-Quran learning. The system can verify the iqra reading. The system will implement voice processing which will match the train data and test data. In general, the system will be divided into two processes, namely voice feature extraction and data matching. This system will use Mel Frequency Cepstrum Coefficients (MFCC) and Dynamic Time Warping (DTW). Mel Frequency Cepstrum Coefficients is used to perform feature extraction, which is to obtain a parameter and information about the characteristics of a person's voice [2]. Meanwhile, Dynamic Time Warping is used for the matching process between train data and test data.

In previous research by Berhaningtyas, Amri and Suprayogi, Hijaiyah speech recognition has been carried out, but in this research developed for the case of the Iqra learning method and its matching was carried out using dynamic time warping [3]. In previous research also by Yuliantari, Hidayat and Wahyunggoro, about feature extraction and recognition of Indonesian vowel speech using Discrete Wavelet Transform for feature extraction and Dynamic Time Warping to match the data [4]. In this
research, the recognition results showed the best accuracy of 80%. The iqra reading verification system with the mel frequency cepstrum coefficient and dynamic time warping that will be built in this research is expected to help teachers in the process of learning to read iqra. so that it can speed up the process of learning to read the Al-Quran. Besides, this research is also expected to become a reference in future research.

2. Method

2.1. Data

Data used are primary data in the form of sound with a waveform data type (.wav). The data is recorded using a voice recorder. Data was acquired from an Al-Quran teacher. One reading is recorded 15 times. The reading used is the Iqra book volume 1 page 1. There are ten readings on page 1. So a total of 150 data is obtained, which will be used as training data as many as 100 data and 50 as test data. The Iqra book volume 1 page 1, as shown in Figure 1.

2.2. Flowchart

The system built consists of two processes, namely the training process and the test process. In the training process, the system input data is a voice signal with wav data type. First, the data will go through a pre-processing process. In this process, noise in signal data will be reduced. After going through pre-processing, the next step is the feature extraction process with the mel frequency cepstrum coefficient. The result of the feature extraction process is a matrix-vector feature with a frame size x coefficient, where the frame is the number of frames formed, and the coefficient is the specified number of MFCC coefficients. The next process is testing. System input data is a voice signal with wav data type. In the process of testing the data will go through the same process as the training process. The results of the extraction process obtained feature vectors which would then be matched

Figure 1. Iqra book volume 1 page 1 [5]
with the results of training data extraction using dynamic time warping. The result of the matching will be the output of the system, which is a statement of true reading or false reading. Design of the system process can be illustrated in Figure 2.

**Figure 2.** Flowchart system

2.3. *Mel Frequency Cepstrum Coefficients*

Mel Frequency Cepstrum Coefficients (MFCC) is a method used to perform feature extraction, namely to obtain a parameter and information about the characteristics of a signal. MFCC is the most frequently used method for extracting audio signal features, because it uses logarithmic computation following the scope of human hearing. Where the sound signal will be filtered linearly for low frequencies (below 1000 Hz) and logarithmic for high frequencies (above 1000 Hz). So that it can represent sound parameters well [6].

There are eight processes in signal extraction using mel frequency cepstrum coefficient, DC removal, pre emphasize, frame blocking, windowing fast fourier transform, mel frequency warping, discrete cosine transform and cepstral liftering. Diagram block mel frequency cepstrum coefficient can be illustrated in Figure 3.

**Figure 3.** Diagram block MFCC [7]
2.3.1 DC Removal

DC-Removal is useful for normalizing the input data. Namely by calculating the average sound sample rate, then subtracting the value of each sample from the average value [7].

2.3.2 Pre Emphasize

Pre-emphasize is used to reduce the noise ratio in the input data, to improve signal quality. Pre-emphasize also used to balance the spectrum of the sound signal. Pre-emphasize equations can be shown on the equation (1).

\[ y(n) = s(n) - \alpha s(n-1), \quad 0.9 \leq \alpha \leq 1.0 \tag{1} \]

2.3.3 Frame Blocking

Frame blocking is a process for dividing the sample sounds into several frames or slots. The signal must be shared because the sound signal continues to change due to shifting of the articulations of the vocal production organs. The length of the frame is usually used between 10 to 30 milliseconds. In this process, it is generally carried out overlapping by approximately 30% to 50% of the frame length [8]. This aims to avoid losing the sound characteristics at the intersection of each frame. This process can be illustrated in Figure 3.

2.3.4 Windowing

Windowing is the process of weighing each frame to reduce signal discontinuity at the beginning and end of the frame after frame blocking process [2]. The windowing equation can be displayed on the equation (2).

\[ x(n) = x_i(n) \cdot w(n-1) \tag{2} \]

There are many window functions, but the hamming window function is the most frequently used. The equation can be seen on the equation (3).

\[ w(n) = 0.54 - 0.46 \cos \frac{2\pi n}{M-1} \tag{3} \]

2.3.5 Fast Fourier Transform

Fast Fourier transform is a very efficient method for solving discrete Fourier transforms which are widely used for signal analysis purposes such as filtering, correlation analysis, and spectrum analysis [9]. FFT converts a signal in the time domain into a frequency domain, where in the time domain the components in the signal cannot be seen explicitly. Features in the time domain include zero-crossing rate, average energy, and maximum amplitude, whereas the frequency domain includes power spectral analysis [6]. The equation for fast Fourier transform can be displayed on the equation (4).

\[ X(m) = \sum_{n=1}^{N-1} x(n) \left[ \cos \left( \frac{2\pi mn}{N} \right) - j \sin \left( \frac{2\pi mn}{N} \right) \right] \tag{4} \]

2.3.6 Mel Frequency Warping

Mel frequency warping generally uses a filter bank. Filter bank is a form of filter which is done to know the energy size of a specific frequency band in the frequency, but for MFCC purposes, filter bank must be applied in the frequency domain [6]. Mel frequency warping equation can be displayed on the equation (5).

\[ Y(t) = \sum_{i=1}^{N} S[i] H_i[j] \tag{5} \]
2.3.7 Discrete Cosine Transform
Discrete cosine transform is the process of decorating the Mel spectrum to produce a good representation of the sound spectral. The concept of DCT is the same as the inverse Fourier transform. The discrete cosine transform equation can be displayed on the equation (6).

\[ C_n = \sum_{k=1}^{K} (\log S_k) \cos \left[ n(k - \frac{1}{2}) \frac{\pi}{K} \right]; n = 1, 2, 3, ..., K \]  

(6)

The zero coefficient of the DCT will generally be omitted, although it indicates the energy of the signal frame. This is done because based on studies that have been done, the zero coefficient is not reliable for speech recognition [10].

2.3.8 Cepstral liftering
Cepstral liftering is a technique used to minimize the sensitivity of the main MFCC feature extraction process. In the feature extraction process there are several weaknesses, the low order of cepstral coefficients is very sensitive to the spectral slope, while the high order is very sensitive to noise [7]. Cepstral liftering smooths the result spectrum of the main processor so that it can be used better for pattern matching.

2.4. Dynamic Time Warping
One of the problems that are quite complicated in the voice recognition process is the different pronunciation time even though they say the same sentence. This causes the matching process between test data, and training data often does not produce optimal values. To overcome this problem, there is a technique that is quite popular in the development of voice signal processing technology, namely dynamic programming or so-called Dynamic Time Warping (DTW) [11]. Dynamic time warping is intended to accommodate the time difference problem. The time difference process refers to the time during the recording process of the test data with the time on the available reference signal template. The basic principle is to provide a range of steps, and it is used to match tracks that indicate a local match [2].

The steps for calculating the dynamic time warping value between the feature vectors of two signals are as follows:
1. Save both features into separate matrices
   \[ x = [x_1, x_2, ..., x_n] \text{ and } y = [y_1, y_2, ..., y_n] \]  

(7)
2. Calculate the global distance with equation (8).

\[ G_{ij} = L_{ij} + \min \begin{cases} G(i-1,j) \\ G(i,j-1) \\ G(i-1,j-1) \end{cases} \]

(8)

3. Result and Discussion
3.1. Python implementation
Program implementation is run in the Command Line Interface (CLI). System input is a training data file and a test data file in the form of a voice signal, and then the system will provide a true output voice or a false voice. Implementation can be represented in Figure 4.
3.2. System performance results with confusion matrix
Evaluation of the system that has been created is carried out using confusion matrix testing to find the accuracy, precision and recall of the system in verifying Iqra readings. Testing in this study was carried out with 10 Iqra readings. Each reading contains 5 test data which is divided into three correct data and two false readings. The first wrong reading is in the form of a “harokat” error, and the second is an error in reading letters. Tests were carried out using MFCC coefficients totalling 11 and 13. The results of the evaluation can be seen in Table 1.

### Table 1. Test results use a combination of the MFCC coefficient 11 and 13.

| Data | 11 coefficient MFCC | 13 coefficient MFCC |
|------|---------------------|---------------------|
|      | Accuracy  | Precision | Recall | Accuracy  | Precision | Recall |
| 1    | 80%       | 75%        | 100%   | 80%       | 75%        | 100%   |
| 2    | 80%       | 75%        | 100%   | 80%       | 75%        | 100%   |
| 3    | 80%       | 75%        | 100%   | 80%       | 75%        | 100%   |
| 4    | 80%       | 75%        | 100%   | 80%       | 75%        | 100%   |
| 5    | 80%       | 75%        | 100%   | 80%       | 75%        | 100%   |
| 6    | 80%       | 75%        | 100%   | 80%       | 75%        | 100%   |
| 7    | 80%       | 75%        | 100%   | 80%       | 75%        | 100%   |
| 8    | 80%       | 75%        | 100%   | 80%       | 75%        | 100%   |
| 9    | 80%       | 75%        | 100%   | 80%       | 75%        | 100%   |
| 10   | 80%       | 75%        | 100%   | 80%       | 75%        | 100%   |
|      | average   | 80%        | 75%    | 100%      | 82%        | 75.5%  |

3.2.1. Analysis of test results on spoken reading
The system test results of the spoken reading can be shown in Figure 5. It shows the test results using the 13 MFCC coefficients obtained the highest accuracy of 100% in reading 10. The lowest accuracy value was 80% for readings 1-9. The average accuracy obtained was 82%.

![Figure 5. Graph of the test results of the spoken reading](image-url)
3.2.2. Analysis of test results on the number of MFCC coefficients
The results of system testing on the number of MFCC coefficients used can be shown in Figure 6. It shows the average accuracy obtained by testing using the 11 MFCC coefficients was obtained 80% while the use of 13 MFCC coefficients was obtained 82%. From these results, it can be concluded that the greater the number of MFCC coefficients used, the better the system's ability to perform voice recognition. However, the more the number of MFCC coefficients used, the time required for the process will also be longer.

![Figure 6. Graph of test results against the number of MFCC coefficients](image)

4. Conclusion
The application built can recognize Iqra readings with the best average accuracy at 82%, precision at 75.5% and recall 100% using 13 MFCC coefficients. From these results, it can be concluded that the greater the number of MFCC coefficients used, the better the system's ability to perform voice recognition. However, the more the number of MFCC coefficients used, the time required for the process will also be longer. Research can be developed by adding classification methods to make it better at verifying true or false data.

5. Acknowledgments
I express my gratitude for the support of the parties involved and hopefully this research can become a guideline and direction for further research. In connection with the completion of this research, the authors would like to express their gratitude and high appreciation to Putra Pratama for helping to collect data in this study. In the final section, the author would like to thank everyone who has a special interest in this research and provide useful data in this study for user research.

6. References
[1] Subur M 2016 Pembelajaran Efektif Membaca Al-Qur’an Dengan Metode Iqra’ Di Raudhatul Athfal Al Athfal J. Pendidik. Anak 2 59–74
[2] Permana I S, Indrawaty Y and Zulkarnain A 2018 Implementasi Metode Mfcc Dan Dtw Untuk Pengenalan Jenis Suara Pria Dan Wanita MIND J. 3 49–63
[3] Berhaningtyas Hertiana, Muh Khaerul Amri S.P S 2017 Pengenalan Huruf Hijayyah Berbasis Pengolahan Sinyal Suara dengan Metode Mel Cepstrum Frequency Cepstrum Coefficient (MFCC) Momentum 13 49–52
[4] Yuliantari R V, Hidayat R and Wahyunggoro O 2016 Ekstraksi Ciri Dan Pengenalan Tutur Vokal Bahasa Indonesia Menggunakan Metode Discrete Wavelet Transform (Dwt) Dan Dynamic Time Warping (Dtw) Secara Realtime Proc. Natl. Semin. Sci. Technol. 7th Fac. Eng. Wahid Hasyim Univ. Semarang 173–8
[5] Humam A 2000 Buku Iqro’ Cara Cepat Belajar Membaca Al- Qur’an (Yogyakarta: Balai Litbang LPTQ Nasional)
[6] Suciani F N, Djamal E C, Ilyas R 2018 Identifikasi Nama Surat Juz Amma dengan Perintah Suara Menggunakan MFCC dan Backpropagation proceeding Semin. Nas. Apl. Teknol. Inf. b 18-b 23
[7] Putra D and Resmawan A 2011 Verifikasi Biometrika Suara Menggunakan Metode Mfcc Dan Dtw Lontar Komput. 2 8–21
[8] Utami Y N, Anbaranti N, Teknik F and Telkom U 2015 Perancangan Speaker Recognition Pada Sistem Kendali Lampu Berbasis Mikrokontroler Speaker Recognition Design Based on Lamp Control System Proc. Eng. 2 3332–46

[9] Handoko D T and Kasih P 2018 Destian Tri Handoko Voice Recognition untuk Sistem Keamanan PC Menggunakan Metode MFCC dan DTW Gener. J. 2 57–68

[10] Afrillia Y 2018 Modifikasi Mel-Frequency Cepstral Coefficient (MFCC) Pada Sistem Penghafalan Al-Qur’an Dalam Pengenalan Pola Nagham Al-Qur’an

[11] Furui S 2001 Digital Speech Processing, Synthesis, and Recognition (Basel: Marcel Dekker, Inc.)