Indonesian Journal of Electrical Engineering and Computer Science
Vol. 1, No. 1, January 2016, pp. 168 ~ 175
DOI: 10.11591/ijeecs.v1.i1.pp168-175

Development of Quran Reciter Identification System Using MFCC and Neural Network

Tayseer Mohammed Hasan Asda[1], Teddy Surya Gunawan*[2], Mira Kartiwi[3], Hasmah Mansor[4]
[1,2,4]Department of Electrical and Computer Engineering, International Islamic University Malaysia (IIUM), Malaysia
[2]Visiting Fellow, School of Electrical Engineering and Telecommunications, University of New South Wales (UNSW), Australia
[3]Department of Information Systems, International Islamic University Malaysia (IIUM), Malaysia
*Corresponding author, email: tayseerasda@hotmail.com, tsgunawan@iium.edu.my

Abstract

Currently, the Quran is recited by so many reciters with different ways and voices. Some people like to listen to this reciter and others like to listen to other reciters. Sometimes we hear a very nice recitation of al-Quran and want to know who the reciter is. Therefore, this paper is about the development of Quran reciter recognition and identification system based on mel frequency cepstral coefficient (MFCC) feature extraction and artificial neural network (ANN). From every speech, characteristics from the utterances will be extracted through neural network model. In this paper a database of five Quran reciters is created and used in training and testing. The feature vector will be fed into neural network back propagation learning algorithm for training and identification processes of different speakers. Consequently, 91.2% of the successful match between targets and input occurred with certain number of hidden layers which shows how efficient are mel frequency cepstral coefficient (MFCC) feature extraction and artificial neural network (ANN) in identifying the reciter voice perfectly.

Keywords*: Speaker Identification, Mel Frequency Cepstral Coefficient (MFCC), Neural Network, Vector Quantization, Quran Reciter*

1. Introduction

Speech is an important and essential mean of communication between human beings. It is the unique feature that each and every person has and differs them one from another. Therefore, scientists took an advantage of this unique feature and developed many systems that can recognize and differentiate between one from another by their voices. Today, with the progress noticed in the field of statistical modeling of speech, speech recognition systems have many applications in many fields in our life that require an interacting between machine and human. Speaker recognition is the operation of realizing who speaks depending on what data we have in utterance and sound waves. This operation uses the voice of the one speaking to identify their identity. In many industries significant of activities in recognizing and identifying what people say are being held out, as well as in local institutions and colleges [1]. Intense research activities in this domain have been carried out by various enterprises and universities and have brought up several descents of speaker-identification and recognition systems.

Each person in this world has different and unique voice. Therefore, the Quran recited by so many reciters will absolutely tend to differ a lot from one reciter to another. Though the reciters are reciting the same Quranic sentence, but the way they delivered it will be different. Quran recitation is the reading (tarteel, tajwid, or taghbir) aloud, reciting, or chanting of portions of the Quran. The reciter is called atālī, murattil, mujawwid, or most commonly a qari. Recitation should be done according to rules of pronunciation, intonation, and caesuras established by the Islamic prophet Muhammad (PBUH). Nowadays there are hundreds of reciters reciting Al-Quran and each one of them has his unique voice that differs him from another although he is reading the same surah or ayah [2]. Since the day that Islam was spread all around the world, there raised the need for Quran applications online and audios for lots of famous and trusted reciters reciting the Holy Quran. As the Holy Quran is the Holy book for Muslims which guides them in their lives, they spread it through the media to reach all Muslims around the world. Although it is

Received September 4, 2015; Revised November 16, 2015; Accepted December 2, 2015

a great advantage to spread the Holy Quran through media as to teach Muslims all over the world the real message the Quran carries yet there are people against Islam who are trying to alter and de-format the verses by bringing fake reciters to recite the Quran in the wrong way. Therefore, there comes the need to develop a recognition system that recognize the famous and trusted reciters of the Holy Quran and distinguish them from others.

There are many methods used to identify a speaker identity, but the most three known methods are linear predictive coding (LPC), perceptual linear predictive coefficients (PLP), and mel frequency cepstrum (MFCC). However, the method used in this paper is MFCC because it has many advantages over the other methods, such as, the accuracy. MFCC has high recognition accuracy while LPC and PLP have less recognition accuracy. Furthermore, MFCC is less Complex and more to mimic the human auditory system and it is the most method suitable for Arabic language [3]. Moreover, the MFCC is the most prevalent and dominant method used to extract spectral features is calculating Mel-Frequency Cepstral Coefficients (MFCC). MFCCs are one of the most popular feature extraction techniques used in speech recognition based on frequency domain using the Mel scale which is based on the human ear scale [4].

There are many researchers conducted researches in voice recognition and identification in general, yet only very few researches conducted on Quran reciter recognition and identification system though it is important. One of the researches conducted was titled "Quranic Verse Recitation Recognition Module for Support in j-QAF Learning" which is mainly to study the performance of the recognition of the Arabic verses of Al-Quran and study most of the methods used in this. Another research [5] was mainly to create a recognition system that will be able to spot the errors online in the recitations carried by the clips using recognition techniques. Lastly, another research describes the challenges and techniques used to build a successful verification system of the Quran verses online and develop a new model of Quran verification using speech recognition techniques [6].

However, these researches that were done so far do not have clear experimental analysis and clear results that show if the methods proposed and used to verify and identify the reciter's voice were suitable and successful. Therefore, this research paper is conducted to cover the experimental part and to complete the other researches done so far for the Quranic identification system. This system is mainly made from a data base of five of famous reciters of Al-Quran Al-Kareem and MATLAB software to extract the features of reciters voices using MFCC, then train them, classify them using neural network classifier and see the performance of the system in recognize their voices and identify their identity.

2. Speaker Identification System

Speaker identification system is the process of automatically recognizing who is speaking by using the speaker-specific information included in speech waves to verify identities being claimed by people accessing systems. A significant number of speaker-recognition researches and activities are being carried out in industries, national laboratories and universities. Several enterprises and universities have carried out intense research activities in this domain and have come up with various generations of speaker-recognition systems [7].

2.1. Study of Speaker Recognition System Using MFCC and Vector Quantization

One of the thesis done is based on identifying an unknown speaker given a set of registered speakers. It is assumed that the unknown speaker to be one of the known speakers and tried to develop a model to which it can best fit into [7]. In the first step of generating the speaker recognition model, feature extraction for the voices is applied using two processes, such as cepstral coefficients extraction, and Mel Frequency Cepstral Coefficients calculation.

Feature extraction is the process of extracting the features of a sound wave signal, which is basically the extraction of the fundamental parameters identifying a speech signal. Though several methods are used to extract features of the sound signal, the MFCC is the most significant and frequent method used in the speaker recognition system [8], and this is the method implemented in the MATLAB program in this thesis. The feature extraction steps are given below:

1. Frame Blocking (Divide signal into frames).
2. Windowing (For each frame obtain the amplitude spectrum).
3. Fast Fourier Transform (FFT) (extract the frequency elements).
4. Mel- Frequency Warping (convert to mel spectrum).

5. Cepstrum (take the discrete cosine transform to find the Cepstrum coefficients).

1. Frame Blocking

Since the vocal tract moves mechanically slowly, speech can be assumed to be a random process with slowly varying properties. Therefore, the speech is divided into overlapping frames of 50ms every 20ms. The speech signal is assumed to be stationary over each frame and this property will prove useful in the following steps.

2. Windowing

Each frame has been windowed to increase the correlation of the linear predictive coding (LPC) spectral estimates between consecutive frames in order to minimize the discontinuity of the signal at the beginning and end of each frame.

3. Fast Fourier Transform (FFT)

After finishing windowing, the Fast Fourier transform process will take place, where that the Discrete Fourier transform (or Fast Fourier transform) is performed to transform from time domain to frequency domain so that we get the frequency components of a signal [9].

4. Mel frequency wrapping

It is known that human perception of audio signal frequencies does not follow a linear scale. After the Fast Fourier transform is done, the transformed signal is passed through a set of filters specifically band pass filters to be able to simplify the spectrum without significant loss of data. The studies show that the mel- frequency scale is a linearly spaced below 1 kHz and logarithmically spaced for frequencies above that. Hence, a subjective pitch is measured on scale called the mel scale for each and every tone with actual frequency f (Hz).

5. Cepstrum

The last step is to de-correlate the mel logarithmic magnitude spectrum to the mel frequency cepstral coefficients MFCC, and this process called Discrete Cosine Transform. The cepstrum is the inverse Fourier transform of the frequency spectrum of a signal in a time domain.

The features extracted act as a basis for further development of the speaker identification process. Next, feature mapping is applied using Vector Quantization using LBG (VQLBG) algorithm. Then the obtained results were good. MFCCs for each speaker were computed and vector quantized for efficient representation. The code books were generated using LBG algorithm which optimizes the quantization process. VQ distortion between the resultant codebook and MFCCs of an unknown speaker was taken as the basis for determining the speaker's authenticity. Accuracy of 75% was obtained using VQLBG algorithm [10].

2.2. Study of Speaker Recognition System Using Neural Network

After the whole features of the signal have been extracted by MFCC, the resulted features will be trained using the artificial neural network. As stated in [11], neural network is an interconnected group of artificial neurons which uses a mathematical model or computational model for information processing based on the connectionist to computation. In this research, back propagation neural network is used to classify the speech pattern using MFCC features. The neural network that was used has 3 layers which are input layer, hidden layer and output layer and the structure of the back propagation neural network is shown in the Figure 1. Accuracy can reach up to 90% or more depending in how much hidden layers are used.

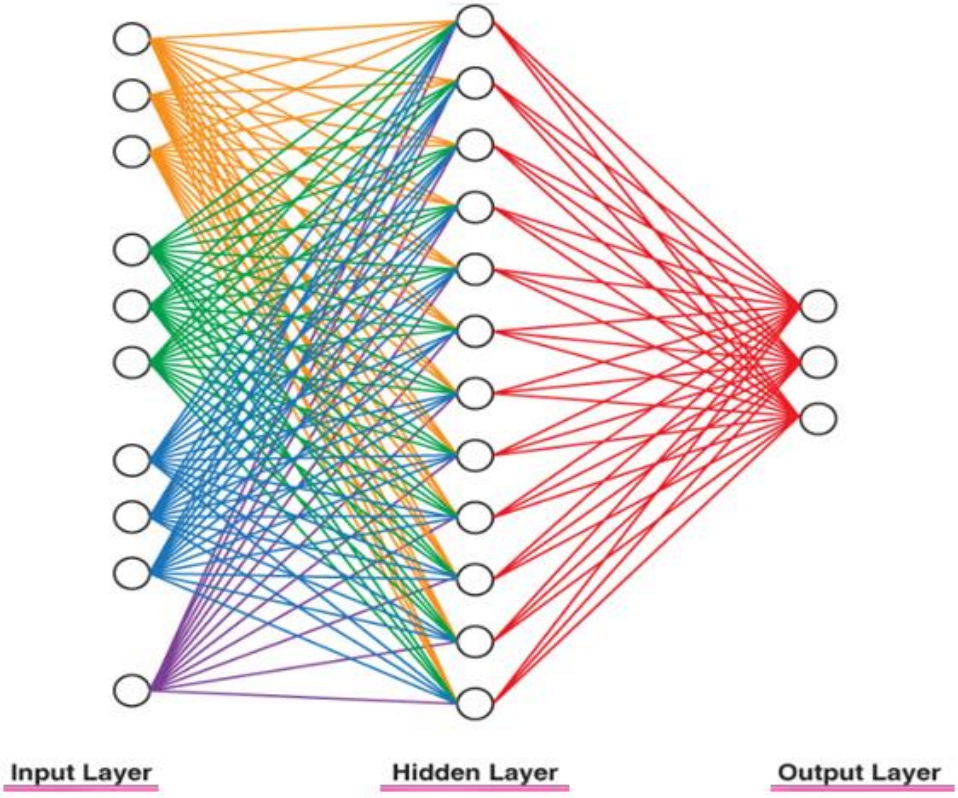

Figure 1. Block Diagram of Neural Network

3. Proposed Quran Reciter Recognition System

Block diagram in Figure 2 shows that the system is divided into two parts. The first part is features extraction and the second part is identification process using neural network.

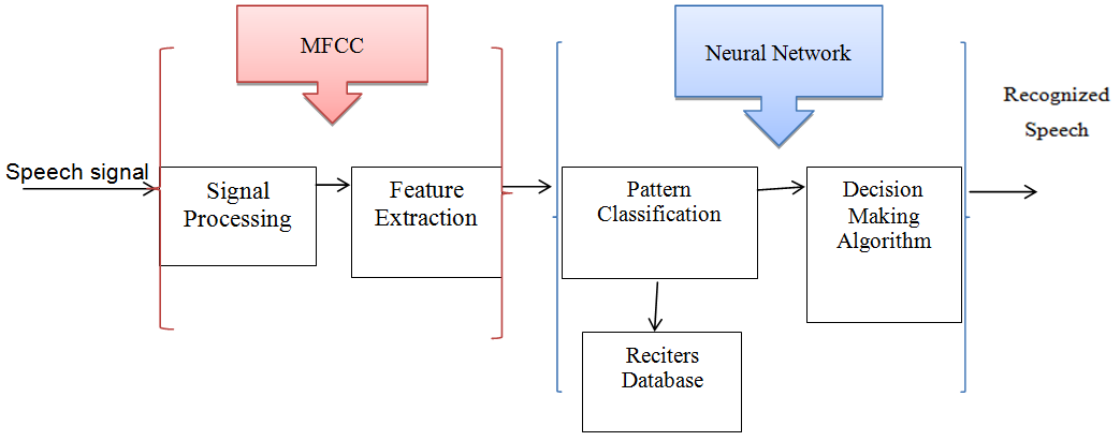

Figure 2. Block Diagram of Reciter Recognition Process

In features extraction, the speech must undergo several process which are pre-emphasis, frame blocking analysis, windowing by using Hammer Windowing and LPC analysis and all of these algorithms are summed in one simple algorithm which is Mel Frequency Cepstral Coefficient (MFCC) features that will extract features from each and every reciter and

the resulted features then will be used as the input for neural network. In the neural network there is a train-validate-test process which is a technique used to reduce model over fitting. The technique is also called early stopping.

One of the major challenges when working with neural networks is a phenomenon called over fitting. Model over fitting occurs when the training algorithm runs too long. The result is a set of values for the weights and biases that generate outputs that almost perfectly match the training data, but when those weights and bias values are used to make predictions on new data, the model has very poor accuracy.

Therefore, the train-validate-test process is designed to help identify when model over fitting starts to occur, so that training can be stopped. Instead of splitting the available data into two sets, train and test, the data is split into three sets: a training set (typically 60 percent of the data), a validation set (20 percent) and a test set (20 percent). The result from the neural network is then being plotted into confusion table to display the percentage of succeeded match features and percentage of mismatch features.

4. Results and Analysis

4.1. Audio Databases

The audio files were downloaded from the internet for a five famous reciters and it was in the form of MP3. Then the MP3 files were converted to WAV files to be read by MATLAB. Lastly, the WAV files were cut to get pure recitation of one reciter with out of any other voices that may interfere with the voice of the reciter.

Table 1. Database for Training

Reciter Name	Surah No.1	Surah No.2	Surah No.3
1- Ahmed Al-ajami	Al-Fatihah	Al-Naba'a	Yasin
2- Abdulrahman Al-sudies	Al-Fatihah	Al-Naba'a	Yasin
3- Nabeel Al-refaey	Al-Fatihah	Al-Naba'a	Yasin
4- Meshari Al-afasi	Al-Fatihah	Al-Naba'a	Yasin
5- Fares Abbad	Al-Fatihah	Al-Naba'a	Yasin

Table 2. Wave Files Details (Training)

Reciter Name	Surah Name	Duration
1- Ahmed Al-ajami	Al-Fatihah	30 seconds
	Al-Naba'a	3 minutes 31 seconds
	Yasin	13 minutes 28 second
2- Abdulrahman Al-sudies	Al-Fatihah	52 seconds
	Al-Naba'a	3 minutes 26 seconds
	Yasin	13 minutes 5 seconds
3- Nabeel Al-refaey	Al-Fatihah	30 seconds
	Al-Naba'a	3 minutes 32 seconds
	Yasin	14 minutes
4- Meshari Al-afasi	Al-Fatihah	40 seconds
	Al-Naba'a	4 minutes 54 seconds
	Yasin	17 minutes 40 second
5- Fares Abbad	Al-Fatihah	37 seconds
	Al-Naba'a	3 minutes 34 seconds
	Yasin	13 minutes 42 second

Table 3. Output Target for Respective Input Reciter

Input Reciter	Output Target Vector
Reciter 1	00001
Reciter 2	00010
Reciter 3	00100
Reciter 4	01000
Reciter 5	10000

4.2 Results Analysis

The neural network modal is fed directly with the output of the MFCC feature extraction with increasing the number of hidden layers from 10 to 20 hidden layers to see the performance of the system. The resulted confusion table is shown in Table 4.

Table 4. Confusion Table Performance for 20 hidden layers

	Reciter 1	Reciter 2	Reciter 3	Reciter 4	Reciter 5	
Reciter 1	20.0%	0.8%	0.7%	0.4%	0.1%	90.9% 9.1%
Reciter 2	0.3%	12.3%	1.2%	0.2%	0.3%	36.0% 14.0%
Reciter 3	0.6%	1.8%	18.3%	1.1%	0.1%	83.8% 16.2%
Reciter 4	0.3%	0.3%	0.8%	18.7%	0.3%	91.4% 8.6%
Reciter 5	0.1%	0.6%	0.4%	0.3%	20.0%	93.4% 6.6%
	93.7% 6.3%	78.1% 21.9%	85.5% 14.5%	90.4% 9.6%	96.0% 4.0%	89.3% 10.7%

TARGET

OUTPUT

From the confusion matrix, the green box (diagonal of the table) shows the successful targets assigned to each class which means the target successfully match with the input while the red box (outside the diagonal) shows the incorrectly targets assigned to class which means mismatch between targets and input. The blue box shows the overall performance according to each process. The green and red percentages show the successful and unsuccessful match percentage. In overall confusion matrix, it combines the training, validating and test process and show the overall successful and unsuccessful percentage. Therefore, after using 20 hidden layers for the processing, the percentage of matching resulted is 89.3%, while the mismatch resulted is 10.7%.

Table 5. Confusion Table Performance for 40 hidden layers

	Reciter 1	Reciter 2	Reciter 3	Reciter 4	Reciter 5	
Reciter 1	20.3%	0.7%	0.5%	0.3%	0.1%	92.8% 7.2%
Reciter 2	0.3%	12.7%	1.1%	0.2%	0.2%	88.0% 12.0%
Reciter 3	0.4%	1.6%	18.9%	0.9%	0.1%	85.9% 14.1%
Reciter 4	0.3%	0.3%	0.7%	19.2%	0.2%	92.7% 7.3%
Reciter 5	0.1%	0.4%	0.4%	0.2%	20.1%	95.6% 4.4%
	95.2% 4.8%	80.8% 19.2%	88.2% 11.8%	92.3% 7.7%	96.7% 3.3%	91.2% 8.8%

TARGET

OUTPUT

From Table 5 it is shown that when the number of hidden layers of the neural network increased from 20 to 40 the total performance of the system got improved by 2%, where the percentage of total matching and recognition of the system becomes 91.2% so that leads to

higher recognition rate. However, when the number of hidden layers further increased to 60 and 80 hidden layers there was no significant change in the total performance. Hence the best performance of the system is when using 40 hidden layers in the neural network as shown in Table 6.

Table 6. performance of the system with different No. of hidden layers

NO. of hidden layers	Time to process	Total Performance
20	10 minutes 59 seconds	89.3% match,10.7% mismatch
40	39 minutes 29 seconds	91.2% match,8.8% mismatch
60	32 minutes 32 seconds	91.2% match,8.8% mismatch
80	1 hour 18 minutes 20 sec	91.1% match,7.9% mismatch

The time of process shown depends on the memory size and the processor speed of the laptop used. Therefore, for this paper the laptop used is Toshiba with a 64 operating system. The installed memory (RAM) is 4 GB. Lastly, the processor used is Intel(R) Pentium(R) CPU P6200 at 2.13 GHz. The result of this research has been represented in terms of percentage to show the performance of the system clearly and how effective it is. However, other research such as [5] which has a similar idea has only shown the overall process and expected results not precise results that were based on an experimental analysis.

5. Conclusion

Overall in this research paper, the purpose is mainly to develop a program in MATLAB that could recognize a reciter's voice as a unique biometric signal and compare it against a database and produce high percentage of matching between the input class and the target class. Knowing that the length of the audio file used has to be more than 10 seconds. Firstly, mel frequency cepstral coefficient (MFCC) features have been used as an input for the neural network and the target of the network has been assigned to the respective input successfully. The percentage of error of the mismatch has successfully reduced by retrain the network using different methods of testing and increases the size of hidden layer. From the overall process, the successful match between targets and input occurred with the 40 hidden layers and with the percentage of 91.2%. This shows that mel frequency cepstral coefficient features can be used in identifying speaker (reciter) perfectly.

References

[1] Hasan MR, Jamil M & Rahman MGRMS. "*Speaker identification using Mel frequency cepstral coefficients. Variations"*. 2004; 1(4).

[2] Ibrahim NJ, Razak Z, Mohd Yusoff Z, Idris MYI, Mohd Tamil E, Mohamed Noor N & Naemah N. "Quranic Verse recitation recognition module for support in J-QAF learning: A Review". *International Journal of Computer Science and Network Security (IJCSNS)*. 2008; 8(8): 207-216.

[3] Kurzekar PK, Deshmukh RR, Waghmare VB & Shrishrimal PP. "*A Comparative Study of Feature Extraction Techniques for Speech Recognition System*". 2014; 3(12).

[4] Dave, N. "Feature extraction methods LPC, PLP and MFCC in speech recognition". *International Journal for Advance Research in Engineering and Technology*. 2013; 1(6): 1-4.

[5] Mohammed A, Sunar MS. "Verification of Quranic Verses in Audio Files using Speech Recognition Techniques". *International Conference of Recent Trends in Information and Communication Technologies (IRICT)*. 2014.

[6] Mohammed A, Sunar MS & Salam MSH. "Quranic Verses Verification using Speech Recognition Techniques". *Jurnal Teknologi*. 2015; *73*(2).

[7] Panda AK. "*Study of speaker recognition systems* ". Doctoral dissertation, National Institute of Technology, Rourkela. 2011.

[8] Janse PV, Magre SB, Kurzekar PK & Deshmukh RR. "A Comparative Study between MFCC and DWT Feature Extraction Technique". In *International Journal of Engineering Research and Technology*. 2014; 3(1).

[9] Sukor A & Syafiq A. "Speaker identification system using MFCC procedure and noise reduction method". Doctoral dissertation, Universiti Tun Hussein Onn Malaysia. 2012.

[10] Kamale HE & Kawitkar RS. "Vector Quantization Approach for Speaker Recognition". *International Journal of Computer Technology and Electronics Engineering*. 2008: 110-114.
[11] Yee CS & Ahmad AM. "*Mel Frequency Cepstral Coefficients for Speaker Recognition Using Gaussian Mixture Model-Artificial Neural Network Mode*" l. *University of Technology Malaysia*.
[12] Narang S & Gupta MD. "*Speech Feature Extraction Techniques: A Review*". 2015.
[13] Hönig F, Stemmer G, Hacker C & Brugnara F. "Revising Perceptual Linear Prediction (PLP)". In *INTERSPEECH*. 2005: 2997-3000.
[14] Fatima N, Wu X & Zheng FT. "Speech unit category based short utterance speaker recognition". *Computer Science and Information Systems*. 2012; 9(4): 1407-1430.
[15] Elminir HK, ElSoud MA & El-Maged LA. "Evaluation of Different Feature Extraction Techniques for Continuous Speech Recognition"*. International Journal of Science and Technology*. 2012; 2(10).
[16] Ellis ED. "Design of a Speaker Recognition Code using Matlab". *Department of Computer and Electrical Engineering–University of Thennessee, Knoxville Tennessee*. 2001; 37996(09).