Journal of Theoretical and Applied Information Technology
15th February 2017. Vol.95. No.3

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

BUILDING AN ASSISTANT MOBILE APPLICATION FOR TEACHING ARABIC PRONUNCIATION USING A NEW APPROACH FOR ARABIC SPEECH RECOGNITION

BASSEL ALKHATIB[1], MOUHAMAD KAWAS[2], AMMAR ALNAHHAS[3], RAMA BONDOK[4], REEM KANNOUS[5]

[1]Assistant Professor at the Faculty of Informatics and Communication Engineering- Arab International University-Syria and the Faculty of Information Technology Engineering-Damascus University. Syria
[2]Teacher assistant at the Faculty of Informatics and Communication Engineering-Arab International University. Syria
[3]Teacher assistant at the Faculty of Informatics and Communication Engineering-Arab International University and the Faculty of Information Technology Engineering-Damascus University.
[4]Fifth Year student at the Faculty of Informatics and Communication Engineering-Arab International University. Syria
[5]Fifth Year student at the Faculty of Informatics and Communication Engineering-Arab International University. Syria
E-mail: [1]b-khateeb@aiu.edu.sy, [2]mouhamadkawas@gmail.com, [3]eng.a.alnahhas@gmail.com
[4]ramabondok@gmail.com, [5]reemkannous.93@hotmail.com.

ABSTRACT

The Arabic language is characterized by its vocal variations. Making its pronunciation a difficult task for Arabic learners. In this paper, we show how we built a mobile application that can detect mispronounced words and guide the user to the correct pronunciation. Foreigners and children can learn Arabic pronunciation in a friendly manner using our application. Our mobile application is customized to help them learn The Holy Quran recitation in particular.
The process of the application compares the user sound (sound signal) of a single word with the set of correct recordings of this word pronunciation. This paper proposes the use of MFCC features to extract features from the speech signal. It also demonstrates the use of a modified version of DTW algorithm to compare the features of the user and the teacher.

Keywords: *Mispronunciation Identification System, Mel-Frequency Cepstrum Coefficients, Dynamic Time Warping, Speech recognition.*

Nomenclature

MFCC	Mel-Frequency Cepstrum Coefficients
DTW	Dynamic Time Warping
ASR	Automatic Speech Recognition

1. INTRODUCTION

Over the last fifty years, speech-processing technology has been growing significantly, due to its potential for a variety of applications in speech recognition, speech correction and speech synthesis [1]. Speech mispronunciation detection is essential for building an assistant system that helps to teach the pronunciation of a specific language [2].

In this paper, a system for mispronunciation identification for the Arabic language is proposed. The Arabic language is the native language for more than 585 million individuals across the world. Making it the third most spoken language [3], the Arabic language has the widest articulatory ladder among all languages [4], i.e.: all of the articulatory organs participate in the creation of sounds from the lips to the glottis. Unlike other languages, that may contain more letters, Arabic sounds are balanced and distinct from each other. These characteristics create a harmony in the Arabic speech.

Self-learning applications using electrical devices like computers or mobile phones are one of the modern learning strategies that are very popular these days; especially language learning application

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

and educational application for acquire speaking a particular language skill. These applications utilize systems that are able to detect errors in the speech of the speakcer; besides, are able to recognize the right spelling. Based on that our work aims at doing a multiple of research and experiments to have a system capable of doing verbal error detection for the person who wants to learn to read the Quran. Quran is a book written in Arabic, and is the most popular in the Arab world; it is noticeable that there are many Quran applications for the young and for the foreigners as well. The goal of this paper can be seen in a number of trends, including modern educational technology currently deployed widely, in applications that help to correct pronunciation for people who have problems with some character pronunciation, and in language learning applications especially in Arabic. It is noted that there are few researches in the field of building systems for detecting verbal mistakes of the Arabic language; in addition, the applications that discover the correct pronunciation of the words of the Quran is very rare. Because of that, we introduce this work to be presented as a mobile application that makes it easier to reach by users and thus will spread larger, and will be useful for many people.

This paper supports the research in the field of audio processing to distinguish incorrect spoken words, which is one of the basic needs for researchers who want to build electrical educational systems, as well as the practice of the method proposed is not only confined for the Arabic language, but also can be applied to any other language.

The proposed system in this paper helps non-native Arabic speakers to learn to recite the holy Quran. It is a mobile application designed for Android devices. The mobile platform was chosen because it can be used easily. The user can pronounce a specific word and the system will process the speech signal to determine if the word was spelled correctly or not. The speech signal is processed by removing the silence, extracting features through the usage of Mel-Frequency Cepstrum Coefficients (MFCC). Finally, the features of the user's and the teacher's (pre-recordings of the word pronounced correctly) are compared using a new proposed modification of Dynamic Time Warping (DTW) algorithm.

The proposed method in this paper suggests a new approach in acoustic signal comparison. DTW algorithm which is used extensively in measuring similarity between two-time series is adjusted to suit the purpose of this research. The algorithm has been modified to measure similarity between two vectors instead of two scalar values. These two vectors represent the acoustic features of a time-window speech of the user of the system and the teacher, correspondingly. This modification has proven its efficiency through the excellent results obtained in the tasted data as shown later.

This paper is organized as the following: section II provides a review of the various methods for building a mispronunciation identification system. In section III, an overview of related work to our study is presented. In section IV, the system architecture is shown. In section V, silence Removal from the speech signal is explained. In section VI, an explanation of the feature extraction algorithm that was used is shown. In section VII, a demonstration of the feature matching algorithm is presented. In section VIII, the stages of building the proposed system are presented. In section IX, the result of the experiment conducted to measure the system performance is shown. Finally, Section X concludes this paper.

2. MISPRONUNCIATION IDENTIFICATION SYSTEM

A Mispronunciation Identification System is a type of assistant teaching systems. It is designed specifically to train the user on the correct pronunciation of words. This kind of system can be built using either ASR (Automatic Speech Recognition) technique or using comparison techniques.

ASR technique is the traditional way to build a mispronunciation identification system [5]. It requires a large amount of training data and a voice lexicon of both the user and the teacher to have all the possible cases of letters' pronunciation. These characteristics delimit the scalability of the teaching system because each new language that needs to be taught, a huge training data and a lexicon for the specified language is required [6].

On the other hand, using a comparison technique to build such a system is very powerful and scalable. This technique uses an algorithm that measures the distance between the user's speech signal and the teacher's speech signal. If the distance between the two signals meets a specific threshold, a deduction is made that the two voices are very near to each other so the user has said the word correctly [7].

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

3. RELATED WORKS

Mispronunciation identification is one of the most important research areas in speech processing. Reference [8] describes a system for the training of Second Language Acquisition Pronunciation (SLAP) for nonnative speakers. In particular, it focuses in helping Chinese students learning the English language. This speech recognition-based system is designed to mimic the valuable interaction between second-language students and a fluent teacher. In this system, when a student speaks a word, the system analyzes it and determines the part of the word that is incorrectly pronounced. A fluent utterance of the word is then played back to the student with emphasis on the mispronounced part of the word. The use of MFCC as a method to extract features from the speech signal has proven its robustness in this system. Also, DTW algorithm is used in the proposed system to measure the distance between the word's pronunciation of the teacher and the student. SLAP system can detect non-native English speakers' mispronunciation, particularly for complicated, multi-syllabic words. However, this system does not target children with learning disabilities. Reference [7] demonstrates using DTW to recognize isolated Arabic words. This system proposes a preprocessing step that includes noise reduction and normalization of the speech signal. Also, Voice Activity Detection (VAD) algorithm is used for detecting the start and end points of voice activity that identifies the silent parts and speech parts of the signal. MFCC approach is adopted in this system due to its effectiveness compared to other well-known feature extraction approaches like Linear Predictive Coding (LPC). Moreover, delta and acceleration coefficients are added to MFCC for the sake of improving the accuracy of Arabic speech recognizer. Finally, DTW is used as a pattern matching algorithm due to its speed and efficiency in detecting similar patterns. This study has expanded the research of automatic speech recognition for Arabic words, which is very limited compared to other languages like English language. The use of voice activity technique has shown a significant impact on system's performance. The results of this research demonstrate a noticeable speech recognition accuracy improvement using MFCC and DTW compared to other HMM and ANN-based approaches. This system achieved a recognition rate of about 98.5%. Reference [9] proposes a system that compares between the signal of the user and the teacher to give features that determine a number of errors by using the distance matrix. The classification process is done using Support Vector Machine (SVM) and Deep Belief Networks (DBNs). This research emphasizes the role of DBN posteriorgrams in improving the relative performance of mispronunciation detection system by at least 10.4%. The study also shows that incorporating non-native data with native data during training would benefit the system. However, the proposed system has only been tested against a small training data set which limits the overall system performance. Reference [10] demonstrates the steps of MFCC to extract features of isolated words of English language. It also takes into consideration the delta energy function to make the feature extraction technique more effective and robust. The delta energy function calculates the time derivatives of (energy + MFCC) which give velocity and acceleration. The outcome of this process is a 39 MFCC feature vector for each frame. This study highlights the use of delta energy coefficients in extracting features from speech signals. However, it limited only to speech identification of isolated English words. Reference [11] presents the implementation of MFCC feature extraction method on Quranic verses. MFCC leads to the conversion of the speech signal into a sequence of acoustic feature vectors. In this system, MFCC features are extended by adding delta or velocity feature and double delta or acceleration feature. The delta features represents the change between frames in the corresponding energy features, while the double delta features represents the change between frames in the corresponding delta features. This feature extending technique yields to a feature vector of 39 values for each frame. The main contribution of this system is to recognize and differentiate the Quranic Arabic utterance and pronunciation based on the feature vectors output produced by using the MFCC feature extraction method. Reference [12] discusses the problem of mistaken recitation of Quranic verses that encounters a lot of Muslims. The authors designed, implemented and tested E-Haifz application that acts like a hafiz expert. E-Hafiz applies MFCC for extracting acoustic feature vectors. Average values of both the user's and the teacher's feature vectors are calculated and then similarity between them is performed by calculating the distance which is the difference between the average values. E-Hafiz is able to facilitate reciting learning of Holy Quran, minimizing errors and mistakes and systematization of the recitation process. The mean recitation ratio of the proposed system is approximately 90%.The main contribution of this research is that it tackled a big issue in the daily life of Muslims, reciting the

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

holy Quran in fear of mistakes. The downside however, is that the system is tested on small number of Quranic verses. Also, this system currently works offline so it can't point out mistakes the user makes during recitation. Reference [16] provides a comprehensive evaluation of Quran recitation recognition techniques. The survey provides recognition rates and descriptions of test data for the approaches considered between LPC and MFCC in the feature extraction process. Focusing on Quran Arabic recitation recognition, it incorporates background on the area, discussion of the techniques, and potential research directions. The result obtained, shows that LPC is the best performance for recognizing the Arabic alphabets of Quran with 50 hidden units of the Recurrent Neural Network with Back-propagation Through Time (99.3%). But, MFCC is still the most popular feature set with 50 hidden units (98.6%), which is computed on a warped frequency scale based on known human auditory perception. The purpose of this research is to upgrade the people's knowledge and understanding on Arabic's alphabet by using Recurrent Neural Network (RNN) and Backpropagation Through Time (BPTT) learning algorithm. However, the study only concentrates on recognizing Arabic letters. Reference [17] presents a system that acts as means of security measures to reduce cases of fraud and theft due to its use of physical characteristics and traits for the identification of individuals. The system is used as an access control key based on voice identification. The most popular cepstrum based method, MFCC, is used to extract the coefficients of voice features. DTW is used to select the pattern that matches the database and input frame in order to minimize the resulting error between them. However, the system is tested against a very small data set. Reference [18] presents MFCC and DTW as tow voice recognition algorithms which are important in improving the voice recognition performance. This research demonstrates the ability of these techniques to authenticate the particular speaker based on the individual information that is included in the voice signal. The results show that MFCC and DTW can be used effectively for voice recognition purposes. However, the test data set is limited to comparing a speech signal of only two speakers. Reference [19] describes an approach of speech recognition by using Mel-Scale Frequency Cepstral Coefficients (MFCC) extracted from the speech signal of spoken words. Principal Component Analysis (PCA) is employed as the supplement in feature dimensional reduction state, prior to training and testing speech samples via Maximum Likelihood Classifier (ML) and Support Vector Machine (SVM). Based on experimental database of total 40 times of spoken words collected under acoustically controlled room, the MFCC extracted features have shown the significant improvement in recognition rates when training the SVM with more MFCC samples randomly selected form the database, compared with the ML classifier. This research emphasizes on MFCC efficiency performance on training scores that agree with improvement in recognition rates when training words with support vector machine. Reference [20] presents effective and robust feature extraction methods using MFCC and its normalized features for isolated digits recognition in English language. Experimental results shows that, MFCC features give more than 95 percent recognition performance on clean data whereas Cepstral Mean Normalized (CMN) features give good performance over noisy data. Recognition rate is highly improved in case of low signal to noise level using Cepstral Normalization. Recognition rate in both, speaker dependent mode and speaker independent mode is improved despite the presence of white Gaussian noise. These features can be used for real time speech recognition. Finally, Reference [21] presents j-QAF, which is a pilot program that suggests rules and regulations to follow, during recitation. The system is useful for people who already know the correct pronunciation and Holy Quran rules. But, it is not suitable for non-Arabic speakers. Mainly, it is a system to help users know Tajweed rules, pointing out mistakes made during recitation. This review paper presents different techniques used for Quran Arabic verse recitation recognition, pointing out advantages and drawbacks. Four techniques are treated. First, Linear Predictive Coding (LPC) that is not considered as a good method, since LPC reduces high and low order Cepstral coefficients into noise when coefficients are transferred into Cepstral coefficients. Second, Perceptual Linear Prediction (PLP) that is better than LPC, since the spectral features remains smooth within the frequency band in PLP and the spectral scale is non-linear Bark scale. Third, Mel-Frequency Cepstral Coefficient (MFCC) which is based on the frequency domain of Mel scale for human ear scale. MFCC is considered the best technique because behavior of acoustic system remains unchanged during transferring the frequency from liner to non-linear scale. Forth, Spectrographic analysis is used for Arabic language phoneme identification. Arabic phonemes are identified by spectrograms that are

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

represented by distinct bands. The review paper also discusses three training and testing method. The first method is Hidden Markov Model (HMM) in which each word is trained independently to get the best likelihood parameters. The second method is Artificial Neural Network (ANN) which is a mathematical based model that recognizes speech in such a way that a person applies to visualizing, analyzing and characterizing the speech to measure its acoustic features. The third method is Vector Quantization (VQ) that uses a set of fixed prototype by matching input vector against each codeword using distortion measure. The author of this paper recommends MFCC as the best approach for feature extraction and HMM or VQ for training and testing. HMM is used when Arabic language recognition has to perform and VQ for English language.

4. SYSTEM ARCHITECTURE

In this section, an overview of our system architecture is presented. Fig. 1 demonstrates the main blocks of the system. First, the silence in the speech signal is removed based on the amplitude. Then, features are extracted using MFCC algorithm. Finally, a comparison is made between the features of the teacher's speech signal and the user's speech signal using DTW algorithm to determine if the user's pronunciation is correct or not.

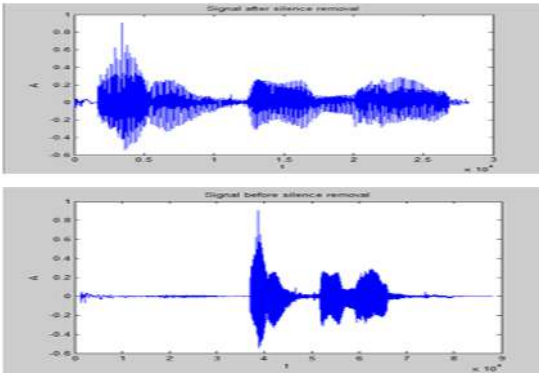

Fig. 1: System Architecture Block Diagram

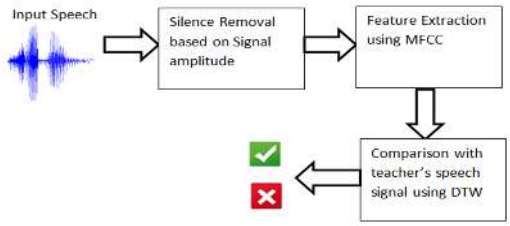

Fig. 2: Signal Silence Removal

5. SILENCE REMOVAL OF SPEECH SIGNAL

This section presents the method used to remove silent parts of the speech signal. Removing silence from the speech signal is the first step in processing the input speech signal of the user. There are various methods to remove silence form the speech signal. The first method is called short-term energy [14]. In this method, we calculate the amount of energy in a speech signal at any time instance. Then, frames that have energy near to zero are discarded. Otherwise, the frame is kept as part of the speech signal. The second method is called ZCR (Zero Crossing Rate) [15]. In the context of discrete-time signals, a zero crossing is said to occur if successive samples have different algebraic signs. The rate at which zero crossings occur is a simple measure of the frequency content of a signal. Zero Crossing Rate is a measure of number of times in a given frame that the amplitude of the speech signal passes through a value of zero. Therefore, the silent part of a speech signal has a high Zero Crossing Rate since its frequency is high [1]. Finally, the third method which is chosen in this research is based on frames' amplitudes. The silent parts of the signal are removed by discarding the signal frames, that their maximum amplitude is smaller than a specific threshold. If the maximum amplitude of the frame is less than 0.3 (this value is determined experimentally) then, the frame is discarded, else it remains a part of the signal. This method proved very good results as show in the following example. Fig. 2 shows the speech signal of the word "الصمد" (Al Sammad) before and after the silence removal process:

6. FEATRURE EXTRACTION

Features extraction starts from an initial set of measured data and builds derived values (features) intended to be informative and non-redundant. Any task that needs to be performed on the original set of data can be done on the features extracted from it; this results in reducing the dimensionalities of the data.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

6.1 Mel-Frequency Cepstrum Coefficients (MFCC)

MFCC is used as the feature of the voice; MFCC is based on human hearing perceptions that cannot perceive frequencies over 1 KHz. In other words, MFCC is based on known variations of the human ear's critical bandwidth with frequency [10].

Fig. 3 shows the steps of MFCC extraction:

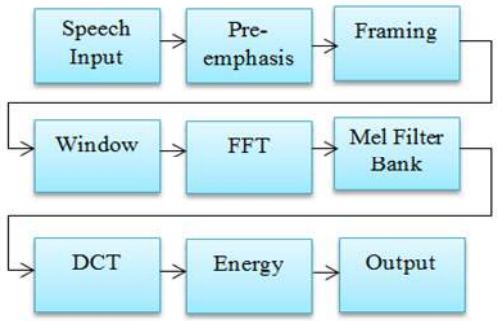

Fig. 3: MFCC Block Diagram

6.1.1 Pre-emphasis

In this step, isolated word sample is passed through a filter which emphasizes higher frequencies. It increases the energy of signal at a higher frequency [10].

6.1.2 Framing

The speech signal is segmented into small duration blocks of 25ms known as frames. The shifting between frames is usually 10ms. Framing required as speech is a time varying signal, but when it is examined over a sufficiently short period of time, short time spectral analysis can be done [10].

6.1.3 Hamming window

Each of these frames is passed to a hamming window function in order to keep the continuity of the signal. The spectral distortion is minimized by using a window to taper the voice sample to zero both at the beginning and at the end of each frame [10].

The following equation demonstrates the use of the hamming window:

$$Y(n) = X(n) \times W(n)$$

Where:

$$W(n) = 0.54 - 0.46\, cos\frac{2\pi n}{N-1}$$

$$0 \leq n \leq N-1$$

Where:

N: Number of samples in each frame

Y (n): Output signal

X (n): Input signal

W (n): Hamming window

Fig. 4 demonstrates the relationship between the speech signal, window size and the amount of shifting the window:

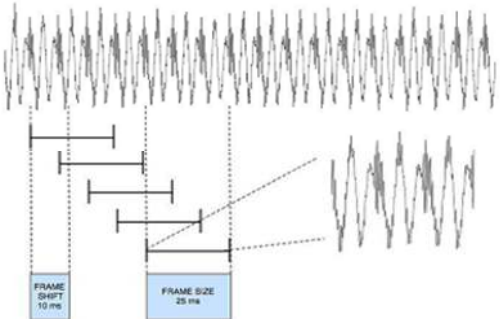

Fig. 4: Window Size and Shift

6.1.4 Fast fourier transform (FFT)

FFT is a process of converting time domain signal into frequency domains. To obtain the magnitude frequency response of each frame, we perform FFT [11].
Fourier transform is given by the equation:

$$X[k] = \sum_{n=0}^{N-1} x[n]e^{-j2\frac{\pi}{N}kn}$$

6.1.5 Mel filter bank

The frequencies range in FFT spectrum is very wide and the voice signal does not follow the linear scale. The bank of filters according to Mel scale is shown in Fig. 5.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

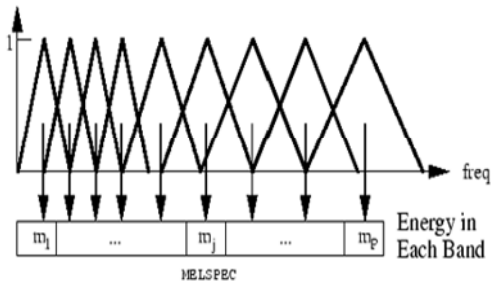

Fig. 5: Mel Filter Bank

Fig. 5 shows a set of triangular filters that are used to compute a weighted sum of filter spectral components so that the output approximate to a Mel scale. Each filter's magnitude frequency response is triangular in shape and equal to unity at the center frequency and decreases linearly to zero at the center frequency of two adjacent filters. Then, each filter output is the sum of its filtered spectral components [13]. The following equation is used to compute the Mel for a given frequency f in Hz:

$$Mel(f) = 2595\, log_{10}\ (1 + \frac{f}{700})$$

6.1.6 Discrete cosine transform (DCT)

This is the process of converting the log Mel spectrum into a time domain using discrete cosine transform (DCT). The result of the conversion is called Mel Frequency Cepstrum Coefficient. The set of this coefficient is called acoustic vectors. Therefore, each input utterance is transformed into a sequence of acoustic vectors [10].

6.1.7 Energy

The energy of each frame is calculated and is added to the acoustic vector [10]. The frame energy is computed by the following equation:

$$Energy = \sum x^2\,[t]$$

The output of the feature extraction process is an acoustic vector of 13 values for each frame. The first value indicates the frame energy and the other twelve values indicate the output of the first six stages of MFCC. In this research we tested records of about fifty Arabic readers using the thirteen values of MFCC and using the twelve values discarding the first value of energy. By experiment, we deduced that discarding the first value of energy is better because the energy of the speech signal differs according to the speaker.

7. FEATURE MATCHING

After the process of feature extraction of each frame of the speech signal, a mechanism to compare extracted features of the user's speech signal against the features of the recordings that were collected (as described fully in the experiment section). By doing so, we can determine if the word pronounced by the user is correct or not.

7.1 Dynamic Time Warping (DTW)

DTW algorithm is based on Dynamic Programming techniques. This algorithm is used to measure the similarity between two-time series which may vary in time or speed. This technique is also used to find the optimal alignment between two-time series if one time series may be warped non-linearly by stretching or shrinking it along its time axis. This warping between the two-time series can be then used to find corresponding regions or to determine the similarity between them [9].

Fig. 6 shows how one time series can be warped to another:

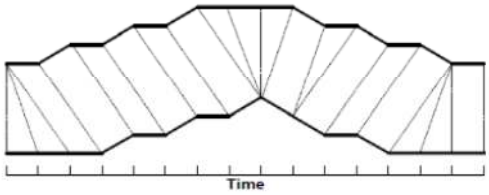

Fig. 6: Dynamic Time Warping between Two Series

Each vertical line connects a point in one time series to its correspondingly similar point in the other time series. The lines have similar values on the y-axis, but have been separated so the vertical lines between them can be viewed more easily. If both of the time series were identical, all of the lines would be straight lines because no warping would be necessary to line up the two-time series. The wrap path distance is a measure of the difference between the two-time series after they have been warped together, which is measured by the sum of the distances between each pair of points connected by the vertical lines as in Fig. 5. Thus, the two-time series that are identical, except for

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

localized stretching of the time axis, will have DTW distance of zero. The aim of DTW is to compare two dynamic patterns and measure its similarity to calculate a minimum distance between them. DTW is computed as described next [7].

Suppose we have two-time series Q and C, of length n and m respectively, where:

$Q= q_1, q_2,\ldots\ldots, q_i, q_n$
$C= c_1, c_2,\ldots\ldots, c_i, c_m$

To align two sequences using DTW, an n-by-m matrix where the (ith, jth) element of the matrix contains the distance d (qi, cj) between the two points qi and cj is constructed. Then, the absolute distance between the values of two sequences is calculated using the Euclidean distance computation:

$$d(qi, cj) = (qi - cj)^2$$

Each matrix element (i, j) corresponds to the alignment between the points qi and cj. Then, accumulated distance is measured by:

$$D(i,j) = \min[\, D(i-1, j-1), D(i-1, j), D(i, j-1)] + d(i,j)$$

The original DTW algorithm measures the distance between scalar values, but this is not the case in our research since we are comparing feature vectors, so we propose a new modification of this algorithm that makes it applicable to vectors rather than scalar values. The first approach is to calculate the Euclidean distance between each two feature vectors. For instance, if we have two vectors q and p then the Euclidean distance between them is given by the following equation:

$$ed(q,p) = \sqrt{\sum_{i=1}^{n} (qi - pi)^2}$$

Where:
N: is the length of the feature vector.

The second approach is to calculate the Cosine similarity between each two feature vectors. The Cosine similarity is a measure of similarity that measures the cosine angle between vectors. The following equation demonstrates the calculation of the Cosine similarity between vector A and B:

$$similarity = \cos(\beta) = \frac{\sum_{i=1}^{n} Ai\ Bi}{\sqrt{\sum_{i=1}^{n} Ai}\sqrt{\sum_{i=1}^{n} Bi}}$$

Where:
N: is the length of the feature vector.
β: is the angle between vector A and B.

The results described in the experiment section demonstrate the Cosine similarity as a better approach than the Euclidean distance approach.

8. SYSTEM DESIGN AND IMPLEMENTATION

Fig. 7 shows the processing steps for a correct pronunciation case. The correct recordings and their features are stored at the server. When the user pronounces the word using the mobile, it is processed using MFCC to extract the features. The features are represented as a two-dimensional array. This array has thirteen rows representing the output vector of MFCC algorithm of the frame. The columns of this array represent the frames of the signal. After features of the user's speech signal have been extracted, the second version of the modified DTW algorithm is applied between the user's features and the features of the recordings stored at the server. Then we take the average of all the values given by the modified DTW algorithm. It is found by experiment that if the final average value is less than a certain threshold then the word is pronounced correctly by the user.

Fig. 8 shows the processing steps for a mispronunciation case. The only difference is that the distance given by the modified DTW algorithm is greater than a certain threshold.

Processing of the speech signal is done using Matlab. MFCC is implemented using the PLP and RASTA and MFCC library of Columbia University. Finally, the original version of DTW algorithm is implemented using a function of Mathworks.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

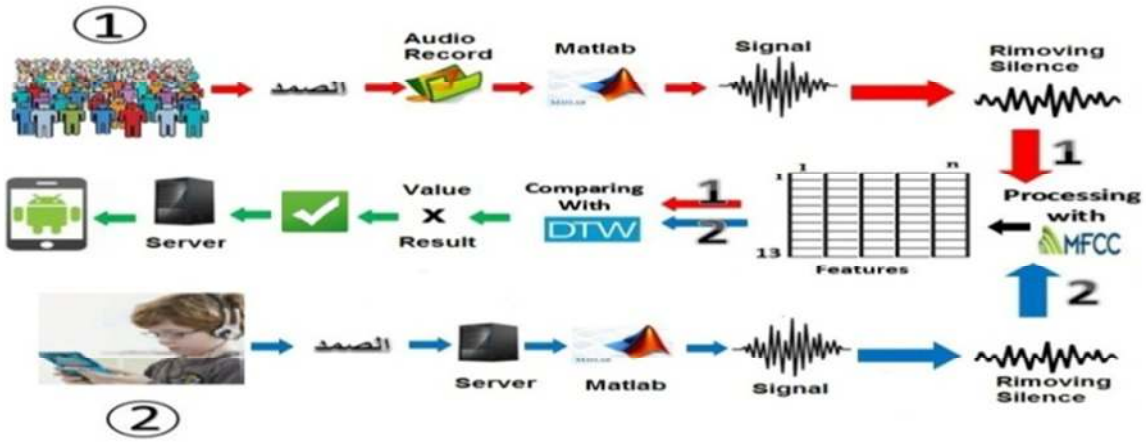

Fig. 7: Correct Pronunciation

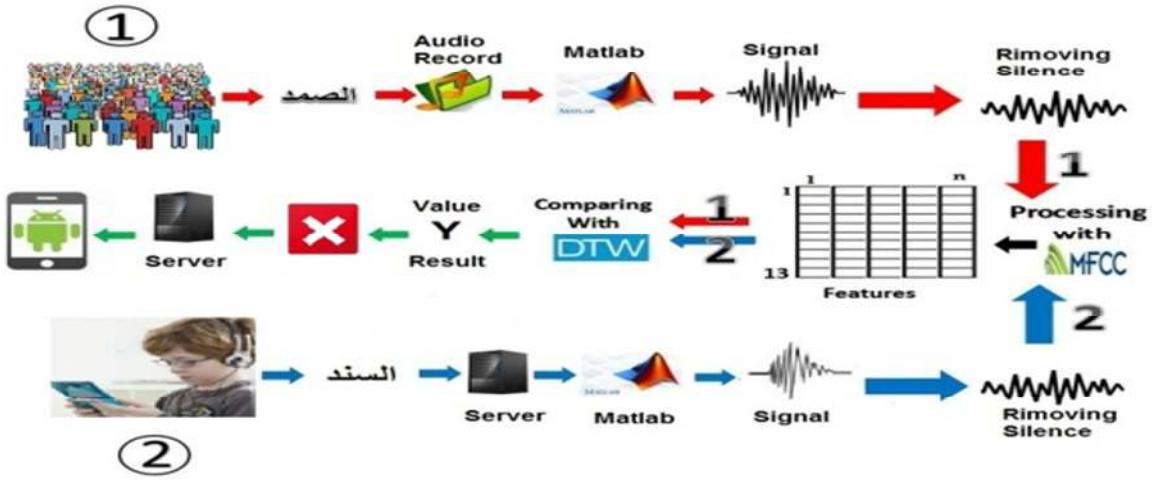

Fig. 8: Wrong Pronunciation

9. EXPERIMENTAL RESULTS

At the beginning, we made our first experiment as follow:

We chose multiple Arabic words that contain different diverse letters; we made sure that they include almost all Arabic letters, this is important so that we can make sure the experiment results can be generalized to cover other words.

We recorded each word from 10 to 15 times, each record is done by a different person from people who are native Arabic speakers; we chose people from different ages (6 to 60 years old) and from both males and females, as it is very important that the training set contains all different sound variations of humans. We used a traditional microphone with normal environmental conditions to make sure no special effect can affect the results; the recording frequency was 44100 Hz.

The records were organized as follows: each record contains one word for specific person, and each word has multiple records for different people with different ages and genders. We noticed that there are small silence periods in the beginning and the end of each record, so that we applied the Removal of silence algorithm to get records without silence part, this can make sure the base samples only contain data of real sound.

We applied MFCC algorithm on all the records, and get the result for each record which is a two dimensional array with constant number of rows equal to 13 rows as mentioned earlier in this paper,

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

and the number of the column of it is change according to the length of the voice of the speaker and the speed of reading the word.

We applied the first edition of DTW algorithm that it used to calculate the distance between two arrays for two similar words with voice of different persons; we noticed that the result is almost a small number close to zero.

Then we applied the first edition of DTW algorithm again to calculate the distance between two arrays for two different words with voice of different persons, and the result was notably much greater than zero.

After conducting a lot of experiments on the records between similar and non-similar words; we realized that when the two words are different the result of DTW will be much bigger than zero ,on the other hand when the two words are similar the DTW result will be much closer the zero.

So that we found that a threshold do exist that can distinguish between similar and non-similar words, that is, if the result of the DTW algorithm is greater than the threshold there is a significant possibility that the two words are different and if it was smaller than the threshold they are likely to be similar.

To find out the precision and recall of this algorithm we did it formally and table-1 shows some of results that we got for some words, it shows how many times the words was recorded and the number of test with other words (with similar word or non-similar word). Where:

True Positive: are the rate of cases where two words were similar and the result of DTW was low.

False Negative: are the rate of cases where two words were not similar and the result of DTW was high.

False Positive: are the rate of cases where two words were similar but the result of DTW was high, and fails to detect the similarity.

True Negative: are the rate of cases where two words were not similar but the result of DTW was low, indicating wrongly that they are similar.

The results of the first experiment were not bad, but they were not that accurate to be implemented in a real system, so we tried to change the DTW algorithm to have better results. We used the cosine distance between MFCC vector of the first word and the MFCC vector of the second word, instead of the Euclidean distance used in the first experiment as mentioned in the last section. Table-2 shows the results after modifying the algorithm it show how the results improved significantly, the columns of the table are the same as Table-1.

The results of the second experiment are much better the first one, so that we used it in our application. We save multiple recordings for each word (10 to 15 records) in the database, each recording is recorded by a different person. Our system receives a recording of a new word from the user, first it removes the silence from the begging and the end of the record; then it calculates the MFCC vector of that recording. Then it compares it using the modified DTW algorithm with all MFCC vectors that we have in the database. If the result of the DTW algorithm were smaller than the error threshold (closer to zero) it means the new word is correct, but if the result was greater than the error Threshold the word is considered not correct.

We applied this approach in a real application, the application contains about 300 words that are the words of the Holy Quran taken from the last 20 versus of it. The system is used to teach children how to correctly pronounce Arabic words of Quran. It shows a good results and supervisors of children are satisfied with the results. We did a survey that asks about the performance of our system and how much it helps improving the teaching process of young children, about 100 teachers participated in the survey, 86 of them say that the application is useful and helped very much.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

TABLE 1: Experiment Results

Word	Number of Records	Number of Testing	Accuracy				Precision	Recall
			True Positive	False Negative	False Positive	True Negative		
قل	10	40	42%	6%	8%	44%	0.84	0.875
هو	10	40	38%	9%	9%	44%	0.81	0.81
الله	10	30	40%	9%	6%	45%	0.86	0.81
أحد	10	30	44%	9%	7%	40%	0.86	0.83
الصمد	15	30	39%	5%	7%	49%	0.84	0.88
يلد	15	40	38%	6%	11%	45%	0.77	0.86
كفوا	15	30	39%	7%	6%	48%	0.86	0.84
يولد	10	40	44%	9%	2%	45%	0.95	0.83

TABLE 2: Experiment Results

Word	Number of Records	Number of Testing	Accuracy				Precision	Recall
			True Positive	False Negative	False Positive	True Negative		
قل	10	40	44%	4%	7%	45%	0.86	0.91
هو	10	40	40%	8%	7%	45%	0.85	0.83
الله	10	30	43%	7%	3%	47%	0.93	0.86
أحد	10	30	47%	6%	6%	41%	0.88	0.88
الصمد	15	30	40%	2%	8%	50%	0.83	0.95
يلد	15	40	40%	7%	7%	46%	0.85	0.85
كفوا	15	30	43%	3%	4%	50%	0.91	0.93
يولد	10	40	45%	6%	2%	47%	0.95	0.88

10. CONCLUSIONS

In this paper, we presented an efficient approach for mispronunciation identification of Quran's words, and we applied a series of steps to distinguish between correct pronunciation and wrong one. Starting from recording voice of the word speaker, then removing silence of this recording, and comparing it with other recording of the word to find if the pronunciation is correct or not. we presented tables and charts showing that the results that we had in these experiments reach our goal to build an intelligent system able to recognize bad pronunciations. The results show that the method we presented produces more accurate output than other methods. The modification we have done to well-known algorithms made significant progress in addressing the spoken words in Arabic to the discovery of the correct pronunciation of the wrong pronunciation.

The excellent results have proved the robustness of MFCC as a feature extraction method and DTW as a feature-matching algorithm.

REFERENCES

[1] Rabiner, L.R., Schafer, R.W. (2000). *Digital Processing of Speech Signals*. New Jersey: Prentice Hall, Inc.

[2] Huang, X., Acero, A., Hon, H. (2001). *Spoken Language Processing*. New Jersey: Prentice Hall, Inc.

[3] List of languages by total number of speakers. (2015). Retrieved from Wikipedia: https://en.wikipedia.org/wiki/List_of_languages_by_total_number_of_speakers.

[4] What is special about the Arabic language. (2012). Retrieved from Lexio Philes: http://www.lexiophiles.com/english/what-is-special-about-the-arabic-language.

[5] Jurafsky, D., Martin, J. H. (1999). *Speech and Language Processing*. New Jersey: Prentice Hall, Inc.

[6] Necibi, K., & Bahi, H. (2012). An arabic mispronunciation detection system by means of automatic speech recognition

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

technology. In *The 13th International Arab Conference on Information Technology Proceedings* (pp. 303-308).

[7] Darabkh, K. A., Khalifeh, A. F., Jafar, I. F., Bathech, B. A., & Sabah, S. W. (2013, May). Efficient DTW-based speech recognition system for isolated words of Arabic language. In *Proceedings of World Academy of Science, Engineering and Technology* (No. 77, p. 689). World Academy of Science, Engineering and Technology (WASET).

[8] Gu, L., & Harris, J. G. (2003, May). SLAP: a system for the detection and correction of pronunciation for second language acquisition. In *Circuits and Systems, 2003. ISCAS'03. Proceedings of the 2003 International Symposium on* (Vol. 2, pp. II-580). IEEE.

[9] Lee, A., Zhang, Y., & Glass, J. (2013, May). Mispronunciation detection via dynamic time warping on deep belief network-based posteriorgrams. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing* (pp. 8227-8231). IEEE.

[10] Singh, P. P., & Rani, P. (2014). An Approach to Extract Feature using MFCC. *International organization of Scientific Research, IOSR Journal of Engineering (IOSRJEN)*, *4*(08), 21-25.

[11] Noor Jamaliah, I., Zaidi, R., Zulkifli, M. Y., Mohd Yamani, I., & Emran, M. T. (2008). Quranic verse recitation feature extraction using Mel-frequency cepstral coefficients (MFCC). In *Proceeding of the 4th International Colloquium on Signal Processing and Its Application (CSPA), Kuala Lumpur, Malaysia* (pp. 13-18).

[12] Muhammad, W. M., Muhammad, R., Muhammad, A., & Martinez-Enriquez, A. M. (2010, November). Voice Content Matching System for Quran Readers. In *Artificial Intelligence (MICAI), 2010 Ninth Mexican International Conference on* (pp. 148-153). IEEE.

[13] Deller, J. R. Jr., Hansen, J. H., Proakis, J. G. (2000) *Discrete Time Processing of Speech Signals, second ed*. New York: IEEE Press.

[14] Greenwood, M., & Kinghorn, A. (1999). SUVing: automatic silence/unvoiced/voiced classification of speech. *Undergraduate Coursework, Department of Computer Science, The University of Sheffield, UK*.

[15] Bachu, R. G., Kopparthi, S., Adapa, B., & Barkana, B. D. (2008). Separation of voiced and unvoiced using zero crossing rate and energy of the speech signal. In *American Society for Engineering Education (ASEE) Zone Conference Proceedings* (pp. 1-7).

[16] Ahmad, A. M., Ismail, S., & Samaon, D. F. (2004, October). Recurrent neural network with backpropagation through time for speech recognition. In *Communications and Information Technology, 2004. ISCIT 2004. IEEE International Symposium on* (Vol. 1, pp. 98-102). IEEE.

[17] Bala, A., Kumar, A., & Birla, N. (2010). Voice command recognition system based on MFCC and DTW. *International Journal of Engineering Science and Technology*, *2*(12), 7335-7342.

[18] Muda, L., Begam, M., & Elamvazuthi, I. (2010). Voice recognition algorithms using mel frequency cepstral coefficient (MFCC) and dynamic time warping (DTW) techniques. *Journal of Computing*, 3(2).

[19] Ittichaichareon, C., Suksri, S., & Yingthawornsuk, T. (2012, July). Speech recognition using MFCC. In *International Conference on Computer Graphics, Simulation and Modeling (ICGSM'2012) July* (pp. 28-29).

[20] Lokhande, N. N., Nehe, N. S., & Vikhe, P. S. (2012, December). MFCC based Robust features for English word Recognition. In *2012 Annual IEEE India Conference (INDICON)* (pp. 798-801). IEEE.

[21] Ibrahim, N. J., Razak, Z., Yusoff, Z. M., Idris, M. Y. I., Tamil, E. M., Noor, N. M., & Rahman, N. N. A. (2008). Quranic verse recitation recognition module for support in j-QAF learning: A review. *International Journal of Computer Science and Network Security (IJCSNS)*, *8*(8).