An efficient speech recognition system for arm-disabled students based on isolated words

Article *in* Computer Applications in Engineering Education · October 2017

DOI: 10.1002/cae.21884

CITATIONS
2

READS
110

6 authors, including:

Khalid Darabkh
University of Jordan
111 PUBLICATIONS **1,421** CITATIONS
SEE PROFILE

Mohammed Hawa
University of Jordan
37 PUBLICATIONS **407** CITATIONS
SEE PROFILE

Ramzi Saifan
University of Jordan
22 PUBLICATIONS **84** CITATIONS
SEE PROFILE

Sharhabeel H. Alnabelsi
Al Ain University of Science and Technology
18 PUBLICATIONS **73** CITATIONS
SEE PROFILE

Received: 14 August 2017 | Accepted: 21 September 2017
DOI: 10.1002/cae.21884

RESEARCH ARTICLE

WILEY

An efficient speech recognition system for arm-disabled students based on isolated words

Khalid A. Darabkh[1] | **Laila Haddad**[1] | **Saadeh Z. Sweidan**[1] | **Mohammed Hawa**[2] | **Ramzi Saifan**[1] | **Sharhabeel H. Alnabelsi**[3]

[1] Department of Computer Engineering, The University of Jordan, Amman, Jordan

[2] Department of Electrical Engineering, The University of Jordan, Amman, Jordan

[3] Department of Computer Engineering, Al-Balqa Applied University, Amman, Jordan

Correspondence
Khalid A. Darabkh, Department of Computer Engineering, The University of Jordan, Amman 11942, Jordan.
Email: k.darabkeh@ju.edu.jo

Abstract

Over the previous decades, a need has emerged to empower human-machine communication systems, which are essential to not only perform actions, but also obtain information especially in education applications. Moreover, any communication system has to introduce an efficient and easy way for interaction with a minimum possible error rate. The keyboard, mouse, trackball, touch-screen, and joystick are all examples of tools which were built to provide mechanical human-to-machine interaction. However, a system with the ability to use oral speech, which is the natural form of communication between humans instead of mechanical communication systems, can be more practical for normal students and even a necessity for arm-disabled students who cannot use their arms to handle traditional education tools like pens and notebooks. In this paper, we present a speech recognition system that allows arm-disabled students to control computers by voice as a helping tool in the educational process. When a student speaks through a microphone, the speech is divided into isolated words which are compared with a predefined database of huge number of spoken words to find a match. After that, each recognized word is translated into its related tasks which will be performed by the computer like opening a teaching application or renaming a file. The speech recognition process discussed in this paper involves two separate approaches; the first approach is based on double thresholds voice activity detection and improved Mel-frequency cepstral coefficients (MFCC), while the second approach is based on discrete wavelet transform along with modified MFCC algorithm. Utilizing the best values for all parameters in just mentioned techniques, our proposed system achieved a recognition rate of 98.7% using the first approach, and 98.86% using the second approach of which is better in ratio than the first one but slower in processing which is a critical point for a real time system. Both proposed approaches were compared with other relevant approaches and their recognition rates were noticeably higher.

KEYWORDS
double thresholds VAD, DTW, educational tools, MFCC, physical disabilities, speech recognition

1 | INTRODUCTION

Computer applications have become an essential part of the educational process in almost every field [5,12,20,22,32, 36,46,47]. Science, mathematics, engineering, computer networking, medical specialties, and even literature are some of the fields which use computer programs in their courses to apply the skills taught theoretically in the class room in a practical way and to create a simulated environment which make the learning process more interactive, realistic, and even efficient compared to the old traditional methods [17,19,21,23–30,33,35,52]. However, to take advantage of these programs, a basic knowledge of computer software is required by the student along with the physical ability to interact with the machine using input and output devices [1,18,31,53,54,64,75].

Human-computer communication field concentrates on providing interaction tools between humans and computers. The mouse, which is an example of a mechanical interactive tool, is one of the most efficient ways for a user to control a computer machine. Human-computer interaction covers mechanical tools, speech recognition, face recognition, and gesture recognition. Each of these areas has its own features and difficulties. However, speech recognition is cost efficient to implement. Moreover, it is a very useful tool for people in different fields of life including educational applications. Actually, besides helping regular students in controlling their computers orally, it can be an essential aid for students with physical disabilities who cannot move their hands or body or even have vision impairments [51].

People who suffer from physical disabilities cannot use standard input devices like keyboard and mouse, so they have only one option for gaining access to the computer (i.e., hands free input methods). In the United States for example, there are over 700,000 people with disabilities of the spinal cord where 70% of them are unemployed [77]. Additionally, many people have other impairments as well, including the 46 million adults in the United States diagnosed with arthritis, the one million with Parkinson's disease, and the 50,000 children and adults with muscular dystrophy [43]. People with such disabilities, especially students, encounter a real challenge to use computers which is an essential requirement for modern life aspects including work, social contact, and even entertainment. Based on that, it is a necessity to enable disabled students to interact with computers in a simple and effective way. Hence, multiple software solutions were proposed with different levels of efficiency and flexibility to create a near normal interaction real time environment between disabled students and computers [7,13,14,34,40,42,67,69]. Accordingly, we were motivated in this work to present a speech recognition system that helps armed-disabled students to communicate with computers for educational applications in a simple yet flexible way and with a high accuracy ratio in implementation.

1.1 | Related works

There are many methodologies, comparable in focus to our proposed work, done to enhance the recognition rate of systems that allow arm-disable students to interact with computers. In reference [69], an Arabic speech recognition system was proposed utilizing open source Carnegie Melon University (CMU) Sphinx-4 and hidden Markov models (HMM). Actually, HMM is considered as an effective statistical model. That means that the system being modeled is assumed to be a Markov process with unknown parameters. Based on this assumption, the main challenge is to determine these hidden parameters, from the observable parameters. The extracted model parameters can then be used to perform further analysis. CMU Sphinx speech recognition system is one of the most robust speech recognizers in English and available for free. This system allows research groups with low budgets to begin conducting research and developing applications more quickly. In reference [40], a pattern matching algorithm based on HMM is implemented using Field Programmable Gate Array (FPGA). Actually, the forward algorithm which is the core of the matching algorithm in HMM was analyzed and modified to be more suitable for FPGA implementations. Moreover, the results showed that the recognition accuracy of the modified algorithm is very close to that in the classical algorithm. However, the FPGA approach brought a gain of achieving higher speed and less occupied area. Even more, the proposed approach can be used for isolated Arabic word recognition and achieved recognition accuracy comparable with the powerful classical recognition systems. On the other hand, Automatic Speech Recognition (ASR) is a technology that allows a computer to identify the words that a person speaks into a microphone or telephone with a wide range of applications. Based on that, investigations were made, in reference [13], between monophone, triphone, syllable, and word-based calculations for recognizing Egyptian Arabic digits. Actually, 39 MFCCs were extracted as features for each recorded voice in the database. MFCC was utilized to train HMMs in which the system matches between the testing word and training database. Speaker-independent HMMs-based speech recognition system was designed using Hidden Markov Model Toolkit (HMMT). The database, used for both training and testing, gathered from 44 Egyptian speakers. The experiments showed that the best recognition performance is obtained when syllables were used to recognize Egyptian Arabic speech compared to the rates obtained for recognition using monophones, triphones, and words.

In reference [14], Arabic numeral recognition strategy was proposed utilizing vector quantization (VQ) and HMM taking into consideration the use of linear predictive cepstral coefficients. Using the vector quantizer, each word is analyzed and represented as a set of acoustical

vectors before being transformed into a symbolic sequence. The training set is composed of 50 speakers where each of them uttered three times the 10 digits. The test set comprises two groups. The first one is composed of 30 speakers who participated in the training stage and the second group comprises 10 other speakers who did not participate at the training set. However, in reference [42], recognizing isolated words in Arabic dialect was done by an investigation of discrete hidden Markov model and dynamic time warping (DTW). In this approach, the authors utilized 13 for not only MFCC coefficients, but also delta and acceleration (delta-delta) coefficients. As a part of an emulated filter-bank made out of 24 triangular weighting functions in Mel-scale, a 256 point FFT was used to find the power spectrum to be ultimately utilized. Subsequently, the natural logarithmic was applied to the 24 filter-bank. As far as the recognition rate is concerned, a frames' overlap length of 512×256 was considered. Additionally, five states were characterized in dynamic HMM-based speech recognizer for every word though transitions between propositions states are conceivable just in left to right course with no states' skipping. No more subtle elements were accounted for these states and transitions. The authors in reference [67] deals with the application of Toeplitz matrices and their minimal Eigen values together with a number of different types of neural networks on speech recognition. The speech signal is looked at as an image and it is treated graphically. The base consists of recorded voices for 20 people from six different countries, not only Arabic. The total number of recorded samples was 5,472 divided into two groups. For each person and voice, five samples were chosen as the test set while the remaining samples are chosen to be the teaching set. The heuristic research of eleven different experiments made on 5,472 samples is summarized and compared with considerable methods of classification and recognition of Arabic spoken-digits (i.e., from 0 to 10). This hybrid way of testing and identification has shown as good results as in written text recognition. Finally, in reference [7], a HMM-based Arabic numeral recognition system was proposed utilizing Discrete Wavelet Transform (DWT) and MFCC. DWT is an efficient tool for decomposing signals into frequency sub-bands. The concept of feature recombination was suggested to be applied to the recognition of spoken Arabic numerals. In details, utterances are decomposed using DWT where the Cepstral coefficients of the resulting sub-bands are calculated. Thenceforth, the obtained coefficients are concatenated to form a single feature vector that is used as an input to the speech classifier or HMM to compute the likelihood. Moreover, simulation results have shown that the recognition rate using the suggested method is comparable with the full-band ASR (conventional) system. However, a drawback of the system is the large dimensionality of the feature vector, which was 48, compared with the full-band system.

In reference [34], an efficient communication system with hearing-impaired people based on isolated words of Arabic language was introduced. Actually, it is a speech recognition system which was developed basically to identify Arabic words that a hearing person speaks into a microphone. Thereafter, the words are translated into a video that implements real Arabic sign language which is easily understood by the hearing-impaired person. The speech recognition process in the system involves mainly voice activity detection (VAD), MFCC, and DTW algorithms. A brief summary of this system is as follows: First, noise reduction and normalization require an advanced preprocessing. After that, VAD algorithm is used to detect speech regions from non-speech regions in the voice signal. In order to facilitate the following tasks, every detected speech regions is divided into multiple manageable segments. The segmentation of speech is divided into two types; the first one divides a sentence into separate words and called "Lexical." The second type, or as called "Phonetic," is based on dividing each word into multiple phones. Moreover, MFCC algorithm, when being compared to linear predictive coding (LPC) and other feature extraction approaches, provides robustness, and effectiveness. Additionally, two coefficients have been added in order to improve the recognition accuracy (delta and acceleration [delta-delta]). Finally, a fast and efficient pattern matching algorithm (i.e., DTW) was employed to detect similar patterns.

1.2 | Our methodology and contributions

In this paper, we propose an efficient system for arm-disabled students that allows them to interact with their computers by just giving vocal commands. Actually, the system can also be very useful for all students not only arm-disabled one. Moreover, besides educational applications, the system is flexible enough to be expanded and applied in many other different fields and to be used by all people. In a brief description, the proposed system, maps the spoken words into tasks that the computer performs. When the system is installed for the first time or a new user is defined, a setup is done to create a database for all the required words used in the system. This is done by recording command words by the new users' voice and extracting their features. Thereupon, the new user can make vocal word commands which their features will be compared with the words features that have already been stored in the database. When a match is found for a specific word, the action or related task to that word is performed like opening a file or creating a new folder. In this process, multiple concepts are involved. For example, a pre-emphasis is made for noise

reduction and normalization when a word is recorded. Moreover, the features extraction for spoken words is done using one of the following two approaches separately:

- First Approach: Double thresholds VAD with MFCCs: The double thresholds VAD are based on short time energy and zero crossing rates which result in having a good performance in case of high signal-to-noise ratio (SNR) [34]. On the other hand, a MFCC algorithm is adopted because it is more robust and effective when compared to other feature extraction approaches [37,55,63].
- Second Approach: Wavelet-based MFCCs: Wavelet transform is applied to the input speech signal that is decomposed into different frequency channels. After that, the MFCCs of the wavelet channels are calculated. The main goal of this approach is to enhance the recognition rate by extracting more features from the speech signal. It is important to mention that delta and acceleration coefficients are added to the MFCCs to enhance the recognition rates in both approaches. Finally, to compare spoken words features with database, pattern matching is applied using DTW algorithm due to its speed and efficiency in detecting patterns [41,65].

The rest of the paper explains the proposed system extensively. Particularly, the description of the system is elaborated in section 2. Sections 3 and 4 present a real-time testing scenario and system performance evaluation, respectively. Finally, section 5 summarizes the work and provides future directions to improve it.

2 | THE PROPOSED SYSTEM

The proposed speech recognition system consists mainly of two phases: database setup and system verification. Both phases are explained in details shortly. The work has been done using two separate approaches which are based on VAD with MFCCs and DWT with MFCCs.

2.1 | Database setup

Database setup is the most critical phase in speech recognition applications. However, Figure 1 explains the process of setting-up the database for a new user or student. At first, each command word is recorded by the voice of the new student. Thereafter, a preprocessing is applied on the samples followed by extracting the feature of the word using either approach of the above. Finally, the features extracted are stored in the database.

The process starts with taking multiple samples for the required word with the voice of a new student or user. As we know, diverse individuals say words differently. This is due to

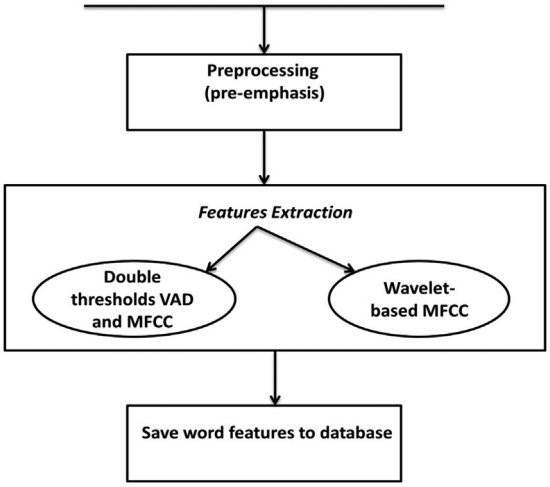

FIGURE 1 Database setup stages

the differences in pitch, slang, and pronunciation between a human and another [37]. A list of required words that are expected to be used by a student while using a computer are recorded in a normal home environment with a sampling frequency of 8,000 Hz and 16-bits depth using a mono channel. An efficient number of samples ought to be taken for every word is 10. On the other hand, preprocessing indicates upgrading some signal qualities to accomplish more precise results through scratching off unsettling influences that may affect the recorded speech quality [63]. Actually, the data signal's spectrum is smoothed by accentuating its high frequency contents. The first order finite impulse response filter is used. The features of each recorded word are extracted using either of the approaches and discussed in details next. The last stage in database setup is to store the extracted features in the database for all required words that were recorded by the new student.

2.1.1 | Features extraction

Two separate approaches are used for the stage of feature extraction. They are detailed as follows:

First approach: Double thresholds VAD with MFCCs

Features extraction in first approach consists of the following two parts:

A) Double Thresholds Voice Activity Detection

One of the real issues that influences the recognizer's productivity is identifying the start and end points of a voice activity [41,65]. VAD is a very widely used technique to improve the performance of such systems. Separation of voiced signal into speech and silence is done on the premise of speech attributes. The signal is cut into contiguous frames. A genuine esteemed non-negative

parameter is connected with every frame [50]. For the time-domain calculations, this parameter is the short-time energy (STE) and used to recognize surd and zero-crossing rate (ZCR) of the frame which basically distinguishes sonant. In the event that this parameter surpasses a certain threshold, the signal frame is named *active*, otherwise it is *inactive* [76]. The double thresholds algorithm sets two thresholds for speech signal. The starting of speech signal is detected by the higher threshold while the lower threshold is used to accurately detect the real starting point of speech signal. The algorithm is described as follows:

- The speech signal is split into multiple frames. The time of each frame is about 30 ms. Actually, there is an overlap of 10–11 ms between adjacent frames. In other words, for a frame of 30 ms or 240 sample points, the overlap takes 10.8 ms or 87 sample points.
- The entire VAD process is divided into four segments: silence segment (status = 0), transition segment (status = 1), speech segment (status = 2) and end segment. In fact, we have characterized a few thresholds for the STE and ZCR as examined later. Moreover, we use the variable *count* as a speech counter, *silence* as a silence counter, and *minlen* as a minimum time threshold as considered in [56].
- The STE of the i^{th} frame is defined as $E_i(n)$ while the ZCR is defined as $Z_i(n)$[56].

$$E_n = \sum_{m=-\infty}^{m} [x(m)w(n-m)]^2 \quad (1)$$

$$Z_n = \sum_{m=-\infty}^{m} |\text{sgn}\,[x(n)] - \text{sgn}\,[x(n-1)]|w(n-m) \quad (2)$$

Initially, we cannot recognize non-speech frames from speech frames. Subsequently, the introductory short-time part is expected to be a non-speech fragment with only a background noise. The threshold of ZCR is IZCT, lower threshold of STE is amp_1 and higher one of STE is amp_2. The threshold values of the first N frames can be computed with N is situated as 5. The threshold of zero-crossing rate is defined as [56,65,74]:

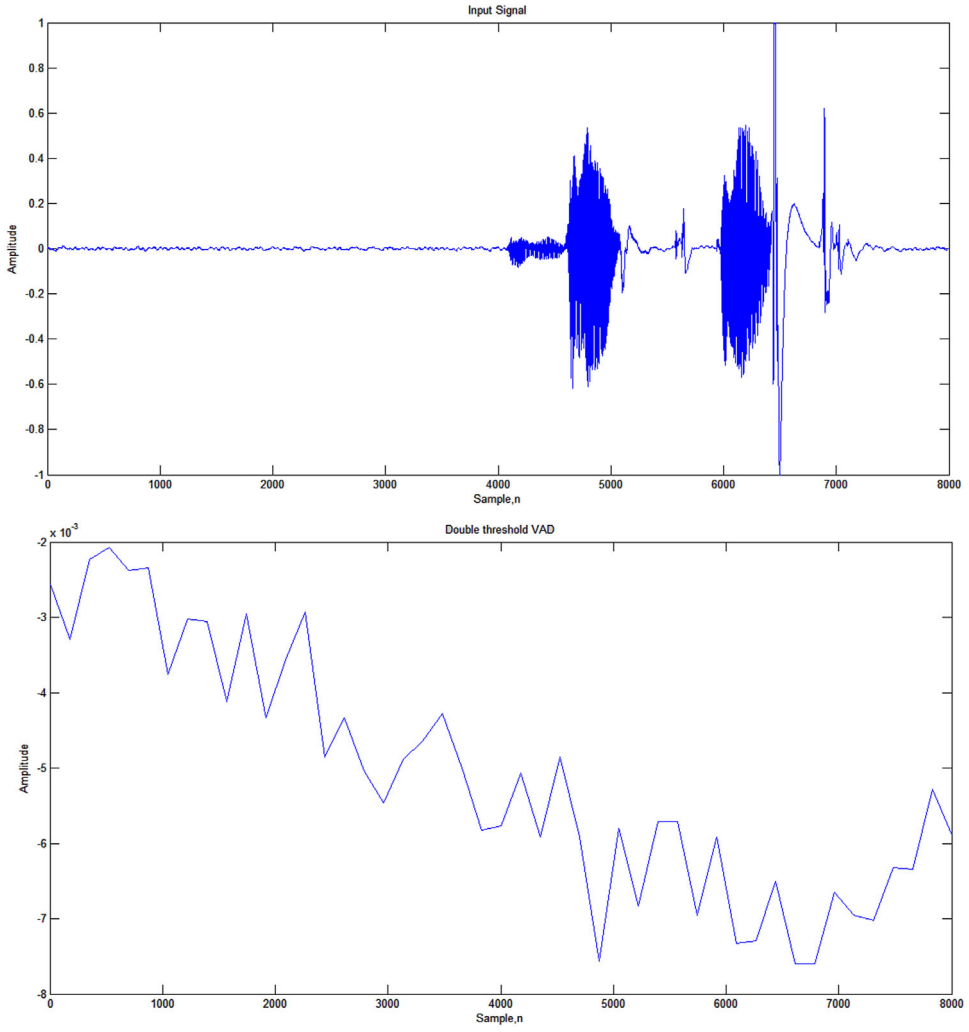

FIGURE 2 The word "desktop" as detected by VAD

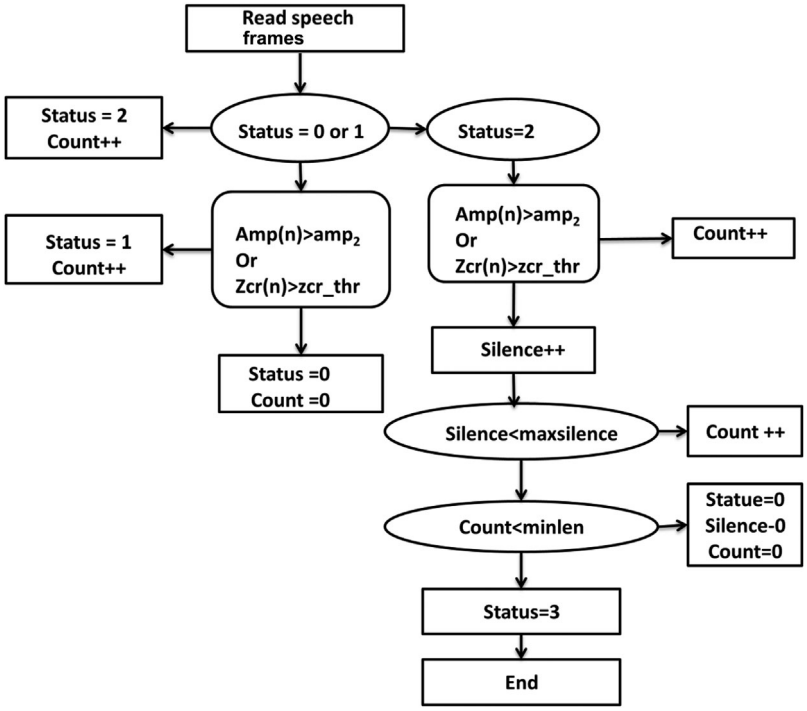

FIGURE 3 VAD based on double thresholds [65,76]

$$\mathrm{IZCT} = \min\left(\mathrm{IF}, \overline{\mathrm{IZC}} + 2\zeta_{\mathrm{IZC}}\right) \tag{3}$$

where IZC is the average of ZCR for the first five frames, ζ_{IZC} is the standard deviation of ZCR, IF is an empirical value and usually sets to 25. Thus, we have

$$\overline{\mathrm{IZC}} = \frac{1}{N}\sum_{i=1}^{N}\mathrm{ZCR}_i(n) \tag{4}$$

$$\zeta_{\mathrm{IZC}} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\mathrm{ZCR}_n(n) - \overline{\mathrm{IZC}}\right)^2} \tag{5}$$

Thus, the STE of the first N frames is calculated. The maximum and minimum among short-time energies of all frames are defined as IMX and IMN, respectively [56,65,68].

Hence, amp_1 and amp_2 are defined as

$$\mathrm{amp}_1 = \min\left(I_1, I_2\right) \tag{6}$$

$$\mathrm{amp}_2 = 5 \times \mathrm{amp}_1 \tag{7}$$

where,

$$I_1 = 0.03 \times (\mathrm{IMX} - \mathrm{IMN}) + \mathrm{IMN} \tag{8}$$

and,

$$I_2 = 4 \times \mathrm{IMN} \tag{9}$$

The output of performing the double thresholds VAD for the word "desktop" is shown in Figure 2. On the other hand, Figure 3 shows a detailed representation for VAD based on double thresholds.

B) Modified Mel-Frequency Cepstrum Coefficients

MFCCs are based on the known variety of the human ear's critical bandwidths with frequency [2]. The work, in this article, is an extension with a modification of replacing inverse discrete cosine transforms by discrete cosine transform (DCT) and accordingly removing the liftering step. The method is based on two sorts of filters, in particular, linearly separated filter and logarithmically divided filter. The phonetically imperative qualities of speech can be caught by representing the signal at the Mel-frequency scale [59]. This scale has a linear frequency separating underneath 1,000 Hz and a logarithmic spacing over 1,000 Hz. Normal speech waveform may differ and rely upon the physical state of speakers' vocal cord. MFCCs are less susceptible to these varieties [61]. The feature extraction is discussed as follows.

Step 1 : Framing

Normally a speech signal is not stationary except from a short-time point of view. This is a result of the fact that the glottal system cannot change immediately. A speech signal typically is stationary in short time windows. Therefore, the signal is divided into frames which correspond to a number of samples [16,34]. However, in this step, every voice signal $x_1(n)$ is parted into J frames. Each

Frame consists of P samples with 36.5% overlapping ratio, so that contiguous frames are divided by T samples (where $T < P$). Moreover, P and T values are 240 and 87 samples, respectively. Consequently, J vectors of length P forms the output signal, which relates to $x_1(p; j)$, where $p = 0,1,2,\ldots,P-1$ and $j = 0,1,2,\ldots,J-1$ [16,34].

Step 2 : Hamming window

In signal processing, a window function (also known as an apodization function or tapering function) is a mathematical function that is zero-valued outside of some chosen interval [45,48,49]. At this point, the output signal is processed through a hamming window which allows diminishing discontinuity of every frame at both ends by applying the following equation [34,62]:

$$\mathrm{Ham}(p) = 0.54 - 0.46\cos\frac{2\pi p}{P-1}, 0 \le p \le P-1 \quad (10)$$

where p represents the sample index and P refers to the frame length in samples. The windowed signal or $x_2(p; j)$, is calculated through applying Ham *(p)* to $x_1(p; j)$ for every frame [34,70].

Step 3 : Fast Fourier Transform (FFT)

FFT algorithms compute the discrete Fourier transform (DFT) of a sequence or its inverse (IFFT) [8–10]. Fourier analysis converts a signal from its original domain (often time or space) to a representation in the frequency domain and vice versa. An FFT computes such transformations rapidly by factorizing the DFT matrix into a product of sparse (mostly zero) factors. Fast Fourier transforms are widely used in many applications in the field of engineering, science, and mathematics. Interestingly, there are many FFT algorithms used in the area mathematics where they range from simple complex-number arithmetic to group theory and number theory [71]. To translate the windowed signal from time domain to a frequency one, *N -point* FFT is used. As a result of that, the characteristics of the speech signal in frequency domain can be easily analyzed. It is noteworthy to mention that the frame length is a power of 2 ($N = 2^p$), and as a result of that, the output signal is $X_2(n; j)$ [34,71].

Step 4 : Mel filter bank

It is widely known that the way that human perception of voice frequencies can be described as nonlinear. In other words, human hearing sensitivity decreases at frequencies higher than 1,000 Hz [50,65]. Intriguingly, a Mel-scale is used to measure different tones where each of them is described by frequency (F) that is measured in hertz, according to following formula [34,65]:

$$F_{\mathrm{mel}} = 2,595\log_{10}\left(1 + \frac{F_{\mathrm{Hz}}}{700}\right) \quad (11)$$

Mel-scale filter bank is a collection of 24 triangular-band-pass filters with both equal spacing before 1 kHz and also logarithmic scale after 1 kHz. Captivatingly, the step, that follows finding the magnitude of $X_2(n; j)$ and using Mel-scale filter, includes finding the Mel spectrum coefficients which is simply performed by summating the filtered results as shown in the following formula [34,65]:

$$\mathrm{Mel}_v = \sum_{n=0}^{N-1}|X_2(n;j)|TF_v^{\mathrm{mel}}(n) \quad (12)$$

where $TF_v^{\mathrm{mel}}(n)$ is the n^{th} triangular filter.

Step 5 : Discrete Cosine Transform

DCT expresses a finite sequence of data points in terms of a sum of cosine functions oscillating at different frequencies. DCTs are important to a numerous number of applications in science and engineering. In particular, a DCT is a Fourier-related transform which is basically similar to the discrete Fourier transform (DFT), but using only real numbers [66,71].

As being formulated in (13), DCT is the preferred technique to use, when planning to get back to the time domain, as of having a highly uncorrelated feature. Actually, logarithm condenses a dynamic range of values whereas the minor differences at high amplitudes do not affect the humans much when compared to that at low amplitudes. According to that, the output magnitude logarithm of Mel-filter bank is calculated first [38,58,66].

$$C(p;j) = \sum_{i=0}^{p-1}\log(\mathrm{Mel}_i)\cos\left(\frac{\pi(2p+1)i}{2P}\right), \quad p = 0,1,\ldots,P-1 \quad (13)$$

Step 6 : Short-time energy

No energy can be captured by cepstral coefficients. Hence, the interesting log feature of signal energy can be used to increase the coefficients that are derived from Mel-cepstrum. Actually, the following energy term is added for every frame [34,58,66]:

$$E_j = \log\sum_{p=0}^{P-1}x_2^2(p;j) \quad (14)$$

Step 7 : Delta and acceleration coefficients

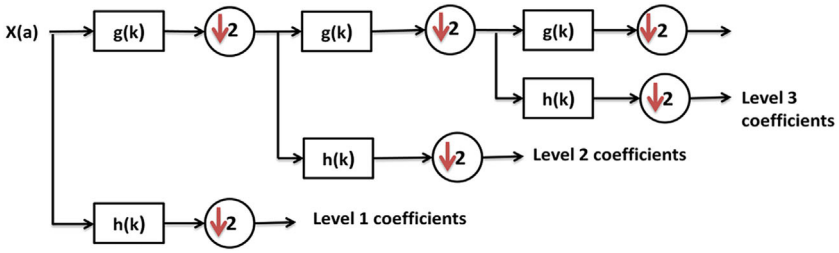

FIGURE 4 Schematic of discrete wavelet decomposition of a speech signal [3,65]

The performance of a speech recognition system can be greatly enhanced by adding time derivatives to the basic static parameters. Naturally, the speech signal is not constant [60]. Based on that, it is worthy to add the changes (i.e., the slopes) in features like delta features and delta acceleration (delta-delta) features. Moreover, it is possible to calculate the delta coefficients by using a linear regression formula considering the size of the regression window as 2C + 1 [4,34,60]:

$$\Delta \mathrm{IC}_l = \frac{\sum_{i=1}^{C} i(\mathrm{IC}_l(m+i) - \mathrm{IC}_l(m-i)}{2\sum_{i=1}^{C} i^2} \tag{15}$$

where IC_l (m) is the m^{th} MFCC coefficient. The delta-delta coefficients are found by utilizing linear regression of delta features. In a nutshell, 39-dimensional features have been utilized including 12 MFCC, 12 delta MFCC features, 1 energy feature, 12 delta-delta MFCC features, 1 delta-delta energy feature, and 1 delta energy feature.

Second approach: Wavelet-based MFCCs

A hybrid feature extraction technique was developed using DWT and MFCC algorithms as follows:

A) Wavelet Transforms

Wavelet transform gives a minimized representation that portrays the energy distribution of the signal in time and frequency domains [39]. We have used DWT to deteriorate the signal into multilevel progressive frequency bands using wavelet functions (Ψ) associated with low and high pass filters. Information caught by wavelet transform depends on the properties of wavelet function family like *Daubechis*, *Symlet*, *Biorthogonal*, *Coiflet,* etc. and on waveform of the target signal [72]. Information

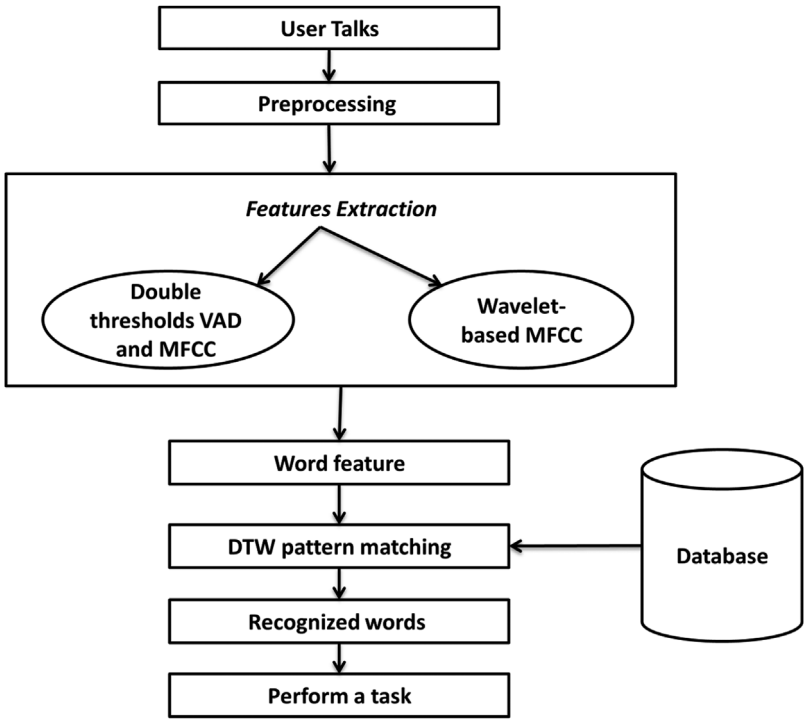

FIGURE 5 System verification stage

extracted by wavelet transforms using a different family of wavelet function is not required to be the same. The wavelet function that gives more useful information, for particular application, is to be picked. Moreover, we have improved precision utilizing *bior3.5* filters [6]. In discrete wavelet decomposition of signal, the output of high pass filter and low pass filter, Y_{high} and Y_{low}, can be represented mathematically by the following [3,78]:

$$Y_{\text{high}}[k] = \sum X[n]g[2k-1] \tag{16}$$

$$Y_{\text{low}}[k] = \sum X[n]h[2k-1] \tag{17}$$

A demonstration of how a 1-dimensional sign is broken into two signs by low-pass and high-pass channels is shown in Figure 4. The down samplers (shown as a down arrow alongside the number 2) eliminate every other sample so that the two remaining signals are approximately half the size of the original. As the figure shows, the low-pass (approximate) signal can be further decomposed, giving a second level of resolution (called an octave). The quantity of conceivable octaves is restricted by the measure of the first signal. As various octaves somewhere around 3 and 6 are common, we utilized 3.

B) Algorithms Fusion

Speech signals are found of two types of information, time and frequency. In time space, sharp variations in signal amplitude are the most significant features. In the frequency domain, in spite of the fact that the dominant frequency channels of speech signals are situated in the middle frequency region, diverse speakers may have distinctive responses in all frequency regions [44]. Along these lines, the conventional techniques which simply consider fixed frequency channels may lose some valuable information in the feature extraction process. Based on the decomposing technique using DWT, one can decompose the speech signal into diverse resolution levels. The characteristic of multiple frequency channels and any change in the smoothness of the signal can then be distinguished to perfectly represent the signals [15].

To this end, the MFCCs are applied to the wavelet channels to extract features characteristics as discussed earlier. MFCCs have the advantage that they can represent sound signals in an efficient way because of the frequency warping property. In this way, the advantages of both techniques are combined in the proposed technique.

2.2 | System verification

After the database has been collected and defined for a specific student by taking multiple spoken samples for all the required words in the system, the system is verified and

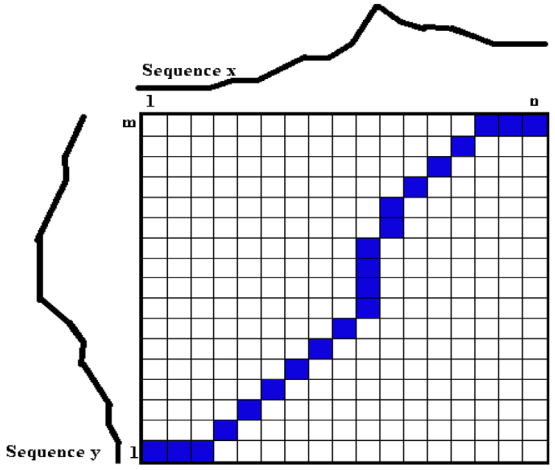

FIGURE 6 Global distance grid

activated for that student. Figure 5 shows the stages for the operation of giving oral commands by the student and translating them into actions or tasks. The operation starts with the student giving vocal commands for the required task. Thereafter, a preprocessing is done in a similar way to the data collection phase. Following that, feature extraction process is executed by one of the approaches discussed earlier. The next step includes a pattern matching process for the spoken words'

FIGURE 7 System avatar

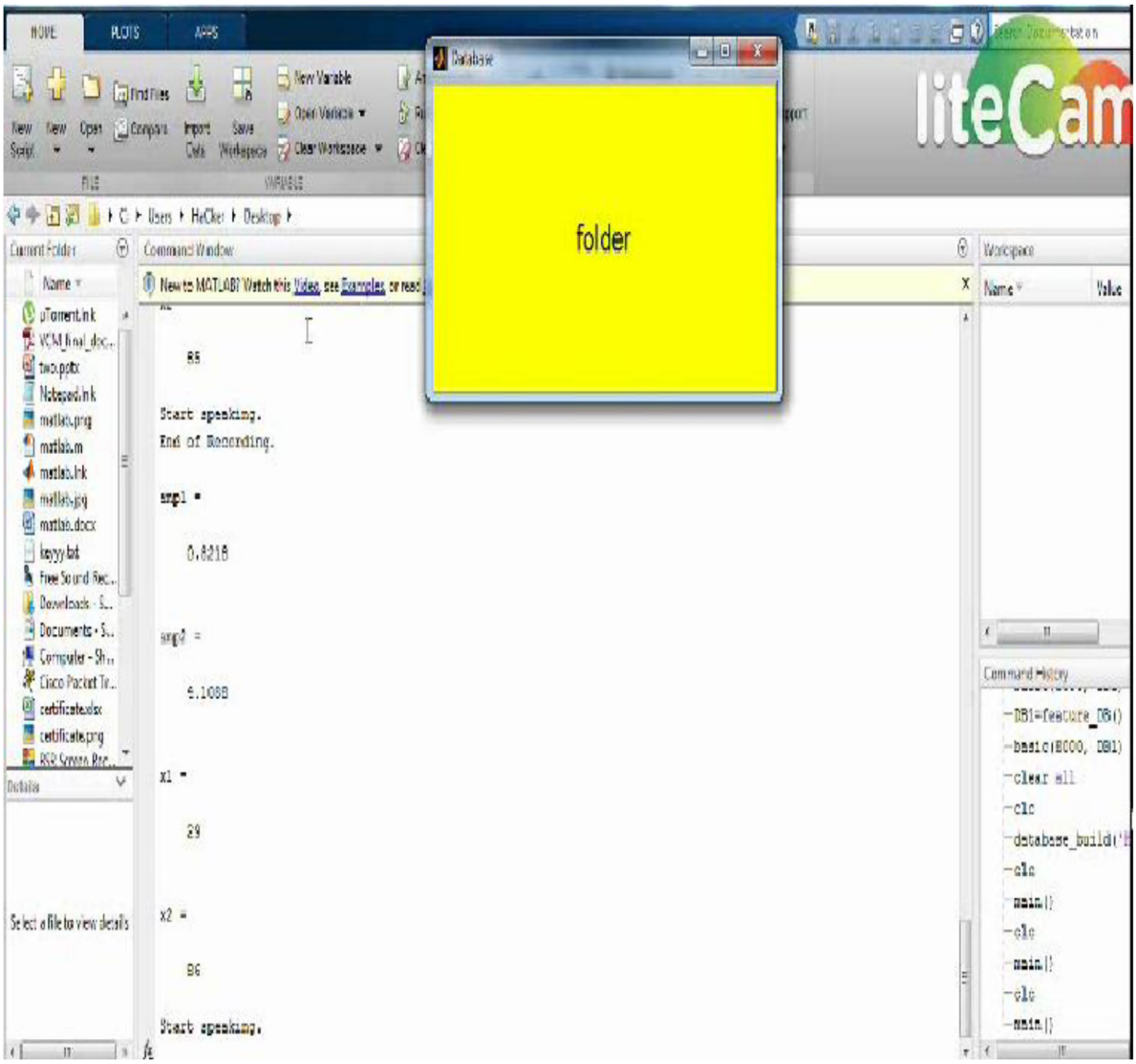

FIGURE 8 Recording the word "folder" to the database

features with the database to find a match. This is done using dynamic time warping which will be explained intensively shortly. Finally, when a match found for a word, the related task is performed.

2.2.1 | Pattern matching

There are many feature-matching techniques used in speech recognizers such as DTW, HMM, and vector quantization [68]. As mentioned earlier, DTW is the used technique in our proposed system. However, the time alignment of different utterances is the core problem for distance measurement in speech recognition. A little shift may drives to an inaccurate recognizable proof. DTW is an effective technique to tackle the time alignment issue [57]. This technique adjusts two groups of highlight vectors by distorting the time pivot redundantly until an ideal match between the two arrangements is found. It performs a piece wise linear mapping of the time axis to align both the signals [73]. Therefore, if we consider two sequences of feature vector in an n-dimensional space as $x \rightarrow [x_1, x_2, \ldots, x_n]$and $y \rightarrow [y_1, y_2, \ldots, y_m]$, then the two sequences will be adjusted on the sides of a grid, with one on the top and other on the left hand side. Both sequences begin on the base left of the grid as shown in Figure 6.

In every cell, a distance measure is set, contrasting the relating components of the two arrangements. The distance between the two points is calculated by means of the Euclidean distance as [38]:

$$\text{Dist}\ (x,\ y) = |x - y|^2$$

$$= \left[(x_1 - y_1)^2 + (x_2 - y_2)^2 + \ldots + (x_m - y_m)^2\right]^{1/2} \quad (18)$$

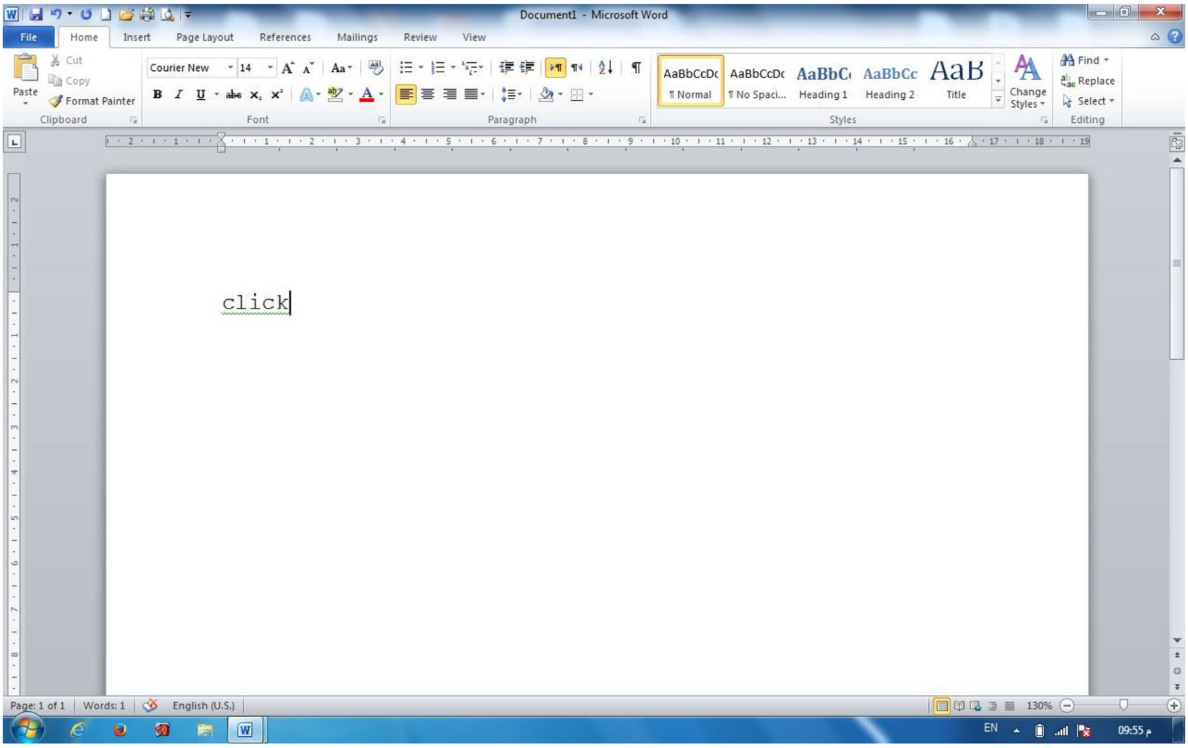

FIGURE 9 Creating a word file and writing on it

The best match between the two sequences is the way through the grid which has the minimum total distance between them, which is termed as global distance. Indeed, it is calculated by finding and going through all the possible routes throughout the grid. For any long sequence, the number of possible paths throughout the grid will be very large. However, the global distance (GD) is obtained using the following recursive formula [38]:

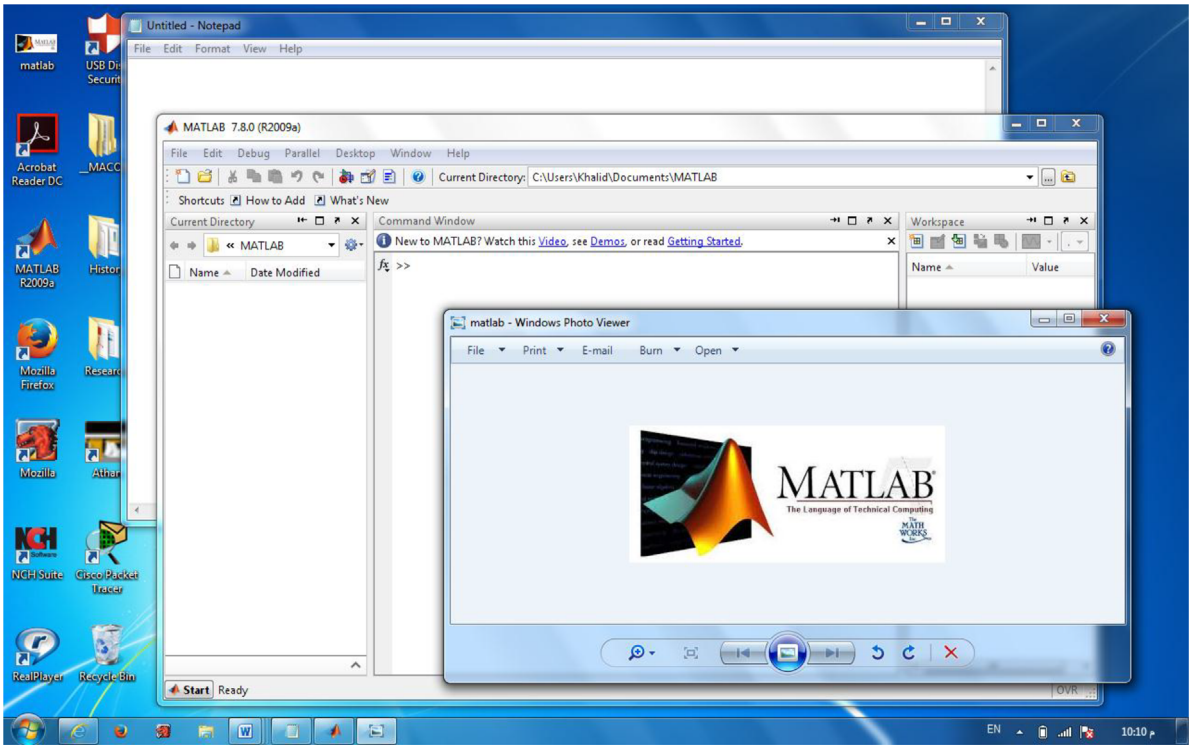

FIGURE 10 Opening an image and a program on desktop

TABLE 1 Recognition rates for chosen tested words

Tested word	Double thresholds VAD + modified MFCC (%)	Proposed algorithm 1: Double thresholds VAD + modified MFCC + Δ + ΔΔ (%)	Proposed algorithm 2: DWT + modified MFCC + Δ + ΔΔ (%)
File	89	91.7	100
Folder	85	96	99
Desktop	100	100	100
Exit	90	100	100
Mouse	90	96.7	98
Keyboard	100	100	100
Start	91	96	100
Web	100	100	100
Browser	85.7	92	99.7
Site	90	95	99
Up	85	91	97.8
Down	85	92.2	98
Left	87	92	100
Right	90	100	100
Click	90	100	100
Undo	87.5	96	100
Save	90	97	99
Copy	90	98.6	100
Notepad	91	100	100
Computer	87	96	99
Picasa	91	95	98
Image	85	93	98
Yes	92	96	100
No	91	97	100
Database	90	97	100
Google	91	96.8	98
Chrome	89	95	99
Firefox	87	97	98
JPG	89	97	98
PNG	91	98.2	99

$$GD_x = LD_x + \text{Min}\left(GD_{x-1,y-1}, GD_{x-1,y}, GD_{x,y-1}\right) \quad (19)$$

where LD refers to the local distance (i.e., Euclidean distance).

3 | REAL-TIME TESTING

In this section, we show a running scenario to preview the phases of the system and show examples of different performed tasks which can be done easily by an arm-disabled student. It is worthy to mention that a human avatar is used to create a friendly human-like communication effect between the student and the system as shown in Figure 7. Interestingly, students in general find it more convenient to interact with a program that has a human face avatar with a calm relaxing voice tone giving the student an imitation of a teacher existence which is much attractive than an automated machined voice.

As formerly mentioned, the student has first to record a group of selected command words used normally during doing home works or running courses simulation programs besides the names of the desktop contents to store their features to the database. For example, the process of recording the sample word "folder" is shown in Figure 8. The avatar tells the student to record 10 samples for each word in a clear voice and environment. This process is done to build a database of the sound features for all the command words and the desktop contents taken by the new user sound.

The verification phase allows the student to perform multiple tasks including, for example, creating, opening, and deleting a file or a folder that include the course files, homework's, and other study documents. They also include opening a website or a browser related to the course official site, controlling the keyboard to print words, writing a report or solving home works, and even controlling the mouse to do right clicks or double clicks in any drag and drop simulation tool. Basically, the commands represent expected actions any

TABLE 2 Recognition accuracy for different works

Speech recognition technique	Achieved accuracy (%)
CMU-based algorithm [69]	85.5
FPGA-based algorithm [40]	91–96 for LPC and 95–98 for MFCC
HMMT-based algorithm [13]	90.75, 92.24, 93.43, and 91.64
VQ-HMM-based algorithm [14]	91
HMM-DTW-based algorithm [42]	86
Heuristic-based algorithm [67]	86.45, 95.82
DWT-based algorithm [7]	61–92, 76–92
VAD-MFCC-based algorithm [34]	95–98
Proposed algorithm 1: Double thresholds VAD with modified MFCC	98.7
Proposed algorithm 2: Wavelet-based modified MFCC	98.86

student does during the use of a computer. Intriguingly, giving commands is done in a form of a dialogue between the avatar and the student where a series of questions and requests are required to be answered to perform a certain operation. The aim is to imitate an intelligent conversation between two people in a humanly environment rather than giving orders to a machine. Figure 9 shows how a word file has been created and the word "click" has been written. The oral dialogue to do this operation is as the following:

- Avatar: what do you want to open please?
- Student: keyboard
- Avatar: do you want to open a word file or other?
- Student: word file
- Avatar: Keyboard is on
- Student: c-l-i-c-k

On the other hand, Figure 10 shows an image called "matlab.jpg" and the notepad program being opened from the desktop as examples of typical student work during studying. This is done by the following oral dialogue:

- Avatar: what do you want to open please?
- Student: image
- Avatar: image's type?
- Student: png
- Avatar: image's name?
- Student: matlab
- Avatar: what do you want to open please?
- Student: program
- Avatar: program's name?
- Student: notepad

4 | PERFORMANCE EVALUATION

To assess the performance of the proposed algorithms, recorded samples, that are stored in the database, were utilized as a part of database setup phase bearing in mind that the base number of tests made to recognize each word was 10. Interestingly, the recognition rate of any word relates to the percentage of the number of correctly recognized words to the number of tested words. The recognition rates for samples of 30 chosen test words are illustrated in Table 1. The rates were calculated using different combinations of features using both proposed approaches.

In the first proposed algorithm, the beneficial outcomes of utilizing double thresholds VAD and modified MFCC on the recognition rate are undeniable. Besides, incorporating delta and acceleration coefficients to the feature set enhances the recognition rate sufficiently which is quite expected as finding delta coefficients requires deriving the first derivative of the feature set which basically means including an imperative parameter as a mirror to the speech changes between consecutive phonemes. It is critical to specify that the first derivative may give noisy results [11]. Therefore, in our proposed algorithm, it is consolidated with utilizing the polynomial approximation approach. Thus, the algorithm's reaction for some tried words was enhanced. Moreover, by fusing the polynomial approximation approach, it would be conceivable to compute the second derivative of the feature (acceleration coefficients) to give more essential information to the features set for the reason of enhancing the overall performance and accuracy of the algorithm. On the other hand, 39 MFCC coefficients were extricated as features for each recorded word in the database where they were utilized to train DTW in which the algorithm matches between the testing word and training database. The acquired recognition rate was 98.7%. As far as the second proposed algorithm is concerned, wavelet-based MFCCs, it has an improved recognition rate in isolated words recognition when compared to MFCCs. In other words, the achieved recognition accuracy was 98.86%. It can be concluded that DWT recognition algorithm combined with modified MFCC has a higher recognition rate for the isolated words recognition. On the other hand, its run time is relatively long when it is compared to the first proposed algorithm. Thus, as a trade-off, both proposed algorithms are interesting and worth presenting. Table 2 summarizes the recognition rates obtained from prior works besides our proposed algorithms. The significance of our proposed algorithms is catching the attention and undeniable. As seen in this table, the recognition rates of prior work are ranged from 61% to 98% which are lower than what are scored in our proposed algorithms without forgetting the fact that all mentioned rates are acquired with the assumption of having clear environment.

5 | CONCLUDING REMARKS AND FUTURE WORK

Computer applications have become very essential in all modern life fields including education. However, it may be a big challenge for disabled students to interact with computer programs. Discovering proficient automatic speech recognition techniques is of great interest for students who use computer programs and even more useful for arm-disabled ones. The proposed system in this article has made the human-machine communication in the educational process easier and more efficient than ever. Moreover, the system presents an easy to use interface with a humanized avatar to create a user friendly environment. Even more, two algorithms have been adopted in this work with a difference of the features

extraction technique. The first approach employs double thresholds VAD technique along with a modified MFCC algorithm. As a matter of fact, it has a positive effect on the system performance. The other approach consists of the modified MFCC algorithm, used in the first approach, consolidated with DWT algorithm. Interestingly, its impact on the system performance is impressive. Moreover, it shows a noticeable strength over the former mechanism. Actually, in view of the time-frequency examination of the wavelet transform, approximations and details resolutions channels are acquired. The MFCCs of the wavelet channels are computed to get the characteristics of the speech signals. Results show that this approach gives better recognition rates than employing just MFCC features. In addition, this technique reduced the problem of noise when dealing with noisy environment. Furthermore, when compared to other methodologies, a discernible speech recognition accuracy improvement is attained. It cannot be missed that incorporating delta and acceleration coefficients into both approaches has helped much in enhancing the overall accuracy of the system. Last, but not least, though the recognition rate of the second approach is higher, the complexity of the first algorithm is lower which means that there is a as a trade-off where both approaches deserved to be presented. As future directions to this work, it will be interesting to enable utterances converted to text to be understood more effectively and at a higher level. Moreover, it will be good to improve our system by adding more features and supporting other languages as well.

REFERENCES

1. G.A. Abandah et al., *Secure national electronic voting system*, J. Inf. Sci. Eng. **30** (2014), 1339–1364.
2. M.I. Abdalla, H.M. Abobakr, and T.S. Gaafar, *DWT and MFCC based feature extraction methods for isolated word recognition*, Int. J. Comput. Appl. **69** (2013), 21–25
3. M.I. Abdalla, and H.S. Ali, *Wavelet-based mel-frequency cepstral coefficients for speaker identification using hidden markov models*, J. Telecommun. **1** (2010), 16–21.
4. N. Aboutabit, D. Beautemps, and L. Besacier, *Automatic identification of vowels in the Cued Speech context*, Proceedings of the International Conference on Auditory-Visual Speech Processing (AVSP 2007), Hilvarenbeek, The Netherlands, 2007, pp. 1–8.
5. A.K. Al-Dhamari, and K.A. Darabkh, *Block-based steganographic algorithm using modulus function and pixel-value differencing*, J. Software Eng. Appl. **10** (2017), 56–77.
6. A.M. Alkababji, *Best wavelet filter for a wavelet neural fricatives recognition system*, AL Rafdain Eng. J. **19** (2011), 138–150
7. W. Alkhaldi, W. Fakhr, and N. Hamdy, *Multi-band based recognition of spoken arabic numerals using wavelet transform*, Proceedings of the 19th National Radio Science Conference (NRSC 2001), Alexandria University, Alexandria, Egypt, 2002, pp. 19–21.
8. R.A. Al Na'mneh, and K.A. Darabkh, *An efficient bit reversal permutation algorithm*, Proceedings of 2013 IEEE International Conference on Robotics, Biomimetics, Intelligent Computational Systems (ROBIONETICS 2013), Yogyakarta, Indonesia, 2013, ppp. 121–124.
9. R.A. Al Na'mneh, and K.A. Darabkh, *New superfast bit reversal algorithms on uniprocessors*, IJCA. **22** (2015), 147–156.
10. R.A. Al Na'mneh, K.A. Darabkh, and I. Jafar, *Efficient bit reversal algorithms in parallel computers*, IJCA. **19** (2012), 154–165
11. M. Al-Zabibi, An acoustic–phonetic approach in automatic arabic speech recognition, Ph.D. Theses, The British Library in Association with UMI, UK, 1990.
12. R. Al-Zubi, K.A. Darabkh, and N. Al-Zubi, *Effect of eyelid and eyelash occlusions on a practical iris recognition system: Analysis and solution*, Int. J. Pattern Recognit Artif Intell. **29** (2015), 1556016-1–1556016-24.
13. M.M. Azmi et al., *Syllable-based automatic arabic speech recognition*, Proceedings of WSEAS International conference of Signal Processing, Robotics and Automation (ISPRA 2008), University of Cambridge, UK, 2008, ppp. 246–250.
14. H. Bahi, and M. Sellami, *Combination of vector quantization and hidden markov models for arabic speech recognition*, Proceedings of the ACS/IEEE International Conference on Computer Systems and Applications (AICCSA 2001), Beirut, Lebanon, 2001, pp. 96–100.
15. U. Bhattacharjee, S. Gogoi, and R. Sharma, *A statistical analysis on the impact of noise on MFCC features for speech recognition*, Proceedings of 2016 International Conference on Recent Advances and Innovations in Engineering (ICRAIE), Jaipur, 2016, pp. 1–5.
16. A. Caplier et al., *Image and video for hearing impaired people*, EURASIP J. Image Video Process. **2007** (2007), 1–14. Article ID 45641, 14 pages.
17. K.A. Darabkh, *Queuing analysis and simulation of wireless access and end point systems using fano decoding*, J. Commun. **5** (2010), 551–561.
18. K.A. Darabkh, *Imperceptible and robust DWT-SVD-based digital audio watermarking algorithm*, J. Software Eng. Appl. **7** (2014), 859–871.
19. K.A. Darabkh, *Fast and upper bounded fano decoding algorithm: Queuing analysis*, T. Emerg. Telecommun. T. **28** (2017), 1–12.
20. K.A. Darabkh, and B. Abu-Jaradeh, *Buffering study over intermediate hops including packet retransmission*, Proceedings of IEEE International Conference on Multimedia Computing and Information Technology (MCIT-2010), Sharjah, U.A.E, 2010, pp. 45–48.
21. K.A. Darabkh, B. Abu-Jaradeh, and I. Jafar, *Incorporating automatic repeat request and thresholds with variable complexity decoding algorithms over wireless networks: Queuing analysis*, IET Commun. **5** (2011), 1377–1393.
22. K.A. Darabkh, A.K. Al-Dhamari, and I.F. Jafar, *A new steganographic algorithm based on multi directional PVD and modified LSB*, Infor. Technol. Control **46** (2017), 16–36.
23. K.A. Darabkh, and O. Alsukour, *Novel protocols for improving the performance of ODMRP and EODMRP over mobile Ad hoc networks*, Int. J. Distrib. Sens. Netw. **2015** (2015), 1–18, Article ID 348967.
24. K.A. Darabkh, R.T. Al-Zubi, and M.T. Jaludi, *New recognition methods for human iris patterns*, Proceedings of 37th IEEE International Convention on Information and Communication

Technology, Electronics and Microelectronics (MIPRO 2014), Opatija, Croatia, 2014, pp. 1187–1191.

25. K.A. Darabkh et al., *An efficient method for feature extraction of human iris patterns*, Proceedings of the 2014 IEEE International Multi-Conference on Systems, Signals & Devices, Conference on Communication & Signal Processing, Castelldefels-Barcelona, Spain, 2014, pp. 1–5.
26. K.A. Darabkh, A.M. Awad, and A.F. Khalifeh, *Efficient PFD-based networking and buffering models for improving video quality over congested links*, Wireless Pers. Commun. **79**, (2014), 293–320.
27. K.A. Darabkh, A.M. Awad, and A.F Khalifeh, *New video discarding policies for improving UDP performance over wired/wireless networks*, Int. J. Network Manage. **25**, (2015), 181–202.
28. K.A. Darabkh, and R.S. Aygun, *Performance evaluation of sequential decoding system for UDP-based systems for wireless multimedia networks*, Proceedings of 2006 International Conference on Wireless Networks (ICWN'06), Las Vegas, Nevada, 2006, ppp. 365–371.
29. K.A. Darabkh, and R. Aygun, *Improving UDP performance using intermediate QoD-aware hop system for wired/wireless multimedia communication systems*, Int. J. Network Manage. **21** (2011), 432–454.
30. K.A. Darabkh et al., *An improved queuing model for packet retransmission policy and variable latency decoders*, IET Commun. **6** (2012), 3315–3328.
31. K.A. Darabkh et al., *An improved image least significant bit replacement method*, In Proceedings of the 37th IEEE International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO), Opatija, Croatia, 2014, pp. 1182–1186.
32. K.A. Darabkh et al., *A new image steganographic approach for secure communication based on LSB replacement method*, Info. Technol. Control **44** (2015), pp. 315–328.
33. K. Darabkh et al., *Efficient DTW-based speech recognition system for isolated words of arabic language*, Proceedings of International Conference on Electrical and Computer Systems Engineering (ICECSE 2013), 2013, pp. 689–692, Lucerne, Switzerland.
34. K.A. Darabkh et al., *A yet efficient communication system with hearing-impaired people based on isolated words of arabic language*, IAENG Inter. J. Comput. Sci. **40** (2013), 183–192.
35. K.A. Darabkh et al., *New arriving process for convolutional codes with adaptive behavior*, Proceedings of IEEE/SSD'12 Multi-conference on Systems, Signals, and Devices, Chemnitz, Germany, 2012, pp. 1–6.
36. K.A. Darabkh, and W.D. Pan, *Stationary queue-size distribution for variable complexity sequential decoders with large timeout*, Proceedings of the 44th ACM Southeast Conference, Melbourne, Florida, 2006, pp. 331–336.
37. S. B. Davis, and P. Mermelstein, *Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences*, IEEE Trans. Acoust. Speech Signal Process. **28** (1980), pp. 357–366.
38. S.D. Dhingra, G. Nijhawan, and P. Pandit, *Isolated speech recognition using MFCC and DTW*, IJAREEIE. **2** (2013), 4085–4092.
39. T. Dutta, *Dynamic time warping based approach to text-dependent speaker identification using spectrograms*, Proceedings of 2008 Congress on Image and Signal Processing, Sanya, China, 2008, pp. 354–360.
40. F.A. Elmisery et al., *A FPGA based HMM for a discrete arabic speech recognition system*, Proceedings of the 15th International Conference on Microelectronics (ICM 2003), Cairo, Egypt, 2003.
41. W.R. Goldmann, and J.R. Mallory, *Overcoming communication barriers: Communicating with deaf people*, Libr. Trends University of Illinois. **41** (1992), 21–30.
42. Z. Hachkar et al., *A Comparison of DHMM and DTW for isolated digits recognition system of arabic language*, Inter. J. Comput. Sci. Eng. **3** (2011), 1002–1008.
43. S. Harada, Harnessing the capacity of the human voice for fluidly controlling computer interface, Ph.D. Dissertation, University of Washington, 2010.
44. C.T. Hsieh, E. Lai, and Y.C. Wang, *Robust speaker identification system based on wavelet transform and Gaussian mixture model*, J. Info. Sci. Eng. **19** (2003), 267–282.
45. I. Jafar, and K.A. Darabkh, *A modified unsharp-masking technique for image contrast enhancement*, Proceedings of IEEE/SSD'11 Multi-conference on Systems, Signals, and Devices, Sousse, Tunisia, 2011, p. 1–6.
46. I. Jafar, K.A. Darabkh, and G. Al-Sukkar, *A rule-based fuzzy inference system for adaptive image contrast enhancement*, Comput. J., **55** (2012), 1041–1057.
47. I. Jafar, K.A. Darabkh, and R. Saifan, *SARDH: A novel sharpening-aware reversible data hiding algorithm*, J. Visual Commun. Image Represent. **39** (2016), 239–252.
48. I. Jafar et al., *An efficient reversible data hiding algorithm using two steganographic images*, Signal Process. **128** (2016), 98–109.
49. I. Jafar, S. Hiary, and K.A. Darabkh, *An improved reversible data hiding algorithm based on modification of prediction errors*, Proceedings of 2014 6th International Conference on Digital Image Processing (ICDIP 2014), SPIE vol. 9159, Athens, Greece, 2014, pp. 91591U-1–91591U-6.
50. B.S. Jinjin Ye, Speech recognition using time domain features from phase space reconstructions, Masters Thesis, Department of Electrical and Computer Engineering, Marquette University, Milwaukee, Wisconsin, 2004.
51. S. Kaur, *Mouse movemment using speech and non speech characteristics of human voice*, Int. J. Eng. Adv. Technol. **1** (2012), 368–374.
52. A. Khalifeh, A. Abbad, and K.A. Darabkh, *An open source TCP/UDP-based network probing tool for real-time packet loss estimation*, Proceedings of 38th IEEE International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO 2015), Opatija, Croatia, 2015, pp. 559–563.
53. A. Khalifeh, K.A. Darabkh, and A. Kamel, *Performance evaluation of voice-controlled online systems*, Proceedings of IEEE/SSD'12 Multi-conference on Systems, Signals, and Devices, Chemnitz, Germany, 2012, pp. 1–6,
54. R. Khalil, A. Khalifeh, and K.A. Darabkh, *Mobile-free driving with android phones: System design and performance evaluation*, Proceedings of IEEE/SSD'12 Multi-conference on Systems, Signals, and Devices, Chemnitz, Germany, 2012, p. 1–6.

55. K. Kirchho et al., Novel Approaches to Arabic Speech Recognition, Technical Report, Johns-Hopkins University, 2002.
56. X. Li, G. Li, and X. Li, Improved Voice Activity Detection based on Iterative Spectral Subtraction and Double Thresholds for CVR, 2008 Workshop on Power Electronics and Intelligent Transportation System, Guangzhou, China, 2008, pp. 153–156.
57. C-T. Lin et al., *EEG-based brain-computer interface for smart living environmental auto-adjustment*, J. Med. Biol. Eng. **30** (2010), pp. 237–245.
58. D. Lodge, *Deaf sentence*, Penguin Books, Wallingford, UK, 2009.
59. P. Melin et al., *Voice recognition with neural networks, type-2 fuzzy logic and genetic algorithms*, Eng. Lett. **13** (2006), 1–9.
60. D. Moores, *Educating the deaf: Psychology, principles, and practices*, Wadsworth Publishing, Lincoln, NE, 2000.
61. L. Muda, M. Begam, and I. Elamvazuthi, *Voice recognition algorithms using mel frequency cepstral coefficient (MFCC) and dynamic time warping (DTW) techniques*, J. Compu. **2** (2010), 138–143.
62. M. Nilsson, and M. Ejnarsson, Speech recognition using hidden markov model (performance evaluation in noisy environment), Masters Thesis, Department of Telecommunications and Signal Processing, Belkinge Institute of Technology, Ronneby, Sweden, 2002.
63. G. Pirelli, (European Commission Joint Research Centre), *The voice project: Giving a voice to the deaf by developing awareness of voice to text recognition capabilities*, Proceedings of the 1998 TIDE Conference, Helsinki, Finland, 1998.
64. E.A. Qaralleh, and K.A. Darabkh, *A new method for teaching microprocessors course using emulation*, Comput. Appl. Eng. Educ. **23** (2015), 455–463.
65. L. Rabiner, and B-H. Juang, *Fundamentals of speech recognition*, Prentice Hall, Upper Saddle River, New Jersey: USA, 1993.
66. S.V. Rezen, and C. Hausman, *Coping with hearing loss: A guide for adults and their families*, Barricade Books Inc., Emeryville, CA, 1993.
67. K. Saeed, and M. Nammous, Heuristic method of Arabic speech recognition, Bialystok University of Technology, Poland, Available online at: http://aragorn.pb.bialystok.pl/~zspinfo/.
68. S. Salvador, and P. Chan, *Toward accurate dynamic time warping in linear time and space*, Intell. Data Anal. **11** (2007), 561–580.
69. H. Satori, M. Harti, and N. Chenfour, Introduction to Arabic Speech Recognition Using CMUSphinx System, Proceedings of Information and Communication Technologies International Symposium (ICTIS'07), Fes, Morocco, pp. 139–115, 2007.
70. V.W. Stern, and M.R. Redden, Selected telecommunications devices for hearing-impaired persons, Office of Technology Assessment, Available online at: http://www.fas.org/ota/reports/8225.pdf, [accessed in August 12, 2017].
71. M.S. Stinson et al., *Deaf and hard-of-hearing students' memory of lectures with speech-to-text and interpreting/note taking services*, J. Spec. Educ. Hammill Institute on Disabilities. **43** (2009), 52–64.
72. Jr. W. C. Stokoe, *Sign language structure: An outline of the visual communication systems of the American deaf*, J. Deaf Stud. Deaf Educ. **10** (2005), pp. 3–37.
73. K. Straetz et al., *An e-learning environment for deaf adults*, Proceedings of the 8th ERCIM Workshop on User Interfaces for All, Vienna, Austria, 2004.
74. X. Sun, A study on efficient robust speech recognition with stochastic dynamic time warping, Available online at: http://eprints.lib.hokudai.ac.jp/dspace/handle/2115/57251.
75. S.Z. Sweadan, and K.A. Darabkh, *A New Efficient Assembly Language Teaching Aid for Intel Processors*, Comput. Appl. Eng. Educ. **23** (2015), 217–238.
76. R. Venkatesha Prasad et al., *Comparison of voice activity detection algorithms for VoIP*, Proceedings of ISCC 2002 Seventh International Symposium on Computers and Communications, Taormina-Giardini Naxos, Italy, 2002, pp. 530–535.
77. S. Venkatraman, and T.V. Padmavathi, *Speech for the disabled*, Proceedings of the 2009 International Multi Conference of Engineers and Computer Scientists (IMECS 2009), vol. 1, Hong Kong, 2009, pp. 567–572.
78. G.K. Verma, U.S. Tiwary, and S. Agrawal, *Multi-algorithm fusion for speech emotion recognition*, Proceedings of International Conference on Advances in Computing and Communications (ACC), Allahabad, India, 2011, pp. 452–459. (Part of the Communications in Computer and Information Science book series [CCIS], volume 192, Springer).

K. A. DARABKH received the PhD degree in Computer Engineering from the University of Alabama in Huntsville, USA, in 2007 with honors. He has joined the Computer Engineering Department at the University of Jordan as an assistant professor since 2007 and has been a tenured full professor since 2016. He is engaged in research mainly on wireless sensor networks, queuing systems and networks, multimedia transmission, and steganography and watermarking. He authored and co-authored of at least 90 research articles and served as a reviewer in many scientific journals and international conferences. Prof. Darabkh is the recipient of 2016 Ali Mango Distinguished Researcher Reward for Scientific Colleges and Research Centers in Jordan. He serves on the Editorial Board of Telecommunication Systems, published by Springer, and Computer Applications in Engineering Education, published by John Wiley & Sons. Additionally, he serves as a TPC member of many reputable IEEE conferences such as GLOBECOM, LCN, VTC-Fall, PIMRC, ISWCS, and IAEAC. Moreover, he is a member of many professional and honorary societies, including Eta Kappa Nu, Tau Beta Pi, Phi Kappa Phi, and Sigma XI. He was selected for inclusion in the Who's Who Among Students in American Universities and Colleges and Marquis Who's Who in the World. As administrative experience at the University of Jordan, he served as an assistant Dean for Computer Affairs in the College of Engineering from September 2008 to September 2010. Additionally, he served as acting head of the Computer Engineering Department from June 2010 to September 2012.

L. HADDAD is a big data Engineer in Amman, Jordan. She got a BSc in Computer Engineering from University of Jordan, and proceeded to work with collaborative engineering professionals since early 2015 in designing and executing solutions for complex business problems involving effectively analyzing and processing terabytes of structured and unstructured data, large scale data warehousing, real-time analytics and reporting solutions.

S. Z. SWEIDAN received his Master degree of Science in Computer Engineering from Jordan University for Science and Technology, Irbid, Jordan in 2007. He is currently a full-time instructor in Computer Engineering Department at the University of Jordan, Amman, Jordan. His research interests include hardware design and implementation, digital arithmetic, cryptography algorithms, and assembly simulators.

M. HAWA graduated from the University of Kansas in 2003 with a PhD degree in Electrical Engineering. He received his MSc degree from University College London in 1999 and his BSc degree from the University of Jordan in 1997. Dr. Hawa is the recipient of the Fulbright Scholarship (1999) and the Shell Centenary Scholarship (1998). He is a published author and a member of the IEEE and IAENG. He is currently a professor of Electrical Engineering at the University of Jordan. His main research interests include: wired and wireless networking, cognitive radio systems, quality-of-service, and peer-to-peer networks.

R. SAIFAN received his BSc and MS degrees in Computer Engineering from Jordan University of Science and Technology, Irbid, Jordan, in 2003 and 2006 respectively, and received his PhD degree in Computer Engineering from Iowa State University in 2012. Since January 2013, he joined as an assistant professor in the Department of Computer Engineering at the University of Jordan. His current research interests include computer networks, computer and network security, cognitive radio networks, and image processing. Saifan published several papers in peer reviewed journals and conferences.

S. H. ALNABELSI is an assistant professor at Computer Engineering Department at Al-Balqa Applied University, Amman, Jordan. He received his PhD in Computer Engineering from Iowa State University, USA, 2012. Also, he received his MSc in Computer Engineering from The University of Alabama in Huntsville, USA, 2007. Dr. Alnabelsi research interests include cognitive radio networks, wireless sensors networks, and network optimization.

How to cite this article: Darabkh KA, Haddad L, Sweidan SZ, Hawa M, Saifan R, Alnabelsi SH. An efficient speech recognition system for arm-disabled students based on isolated words. *Comput Appl Eng Educ*. 2017;1–17. https://doi.org/10.1002/cae.21884