Using Digital Processing of Speech and video to support Interaction between the Deaf Community and Normal People

By

A. E. El-Alfi
Computer Science Department
Mansoura University, Egypt

Y.El-Helaly
Mansoura University,
Egypt

M. M. Ghoniem
Computer Science Department
Mansoura University, Egypt

Research Journal Specific Education
Faculty of Specific Education
Mansoura University

ISSUE NO. 36, OCTOBER. 2014

مجلة بحوث التربية النوعية – جامعة المنصورة
العدد السادس والثلاثون – اكتوبر ٢٠١٤

USING DIGITAL PROCESSING OF SPEECH AND VIDEO TO SUPPORT INTERACTION BETWEEN THE DEAF COMMUNITY AND NORMAL PEOPLE

A. E. El-Alfi* Y.El-Helaly** M. M. Ghoniem***

Abstract

This paper presents a framework for a proposed pattern recognition system to support communication between the deaf people and the normal. The presented system is based on the digital processing of both video and speech. Firstly, it receives the gesture video from the deaf person and converts it to the corresponding spoken word. Secondly, it receives the spoken word from the normal person and converts it to the corresponding gesture. The proposed system comprises the following functions: (1) signal pre-processing; (2) signal analysis; and (3) training and testing. Fig. 1 shows the main block diagram of the proposed system .

Keywords: sign language, Wavelet Packet Decomposition (WPD), Invariant moment, Mel-Frequency Cepstral Coefficients (MFCC), Timbral texture, Weighted Euclidean distance(WED)

1. INTRODUCTION

Human–Computer Interaction (HCI) is getting increasingly important as computer's influence on our lives is becoming more and more significant. With the advancement in the world of computers, the already existing HCI devices (the mouse and the keyboard for example) are not satisfying the increasing demands anymore. Designers are trying to make HCI faster, easier, and more natural. To achieve this, human-to human interaction techniques are being introduced into the field of HCI [1].

One of the most fertile Human-to-Human Interaction fields is the use of hand gestures. People use hand gestures mainly to communicate and to express ideas. The importance of using hand gestures for communication becomes clearer when sign language is considered. The sign language is the fundamental communication method between people who suffer from hearing defects. In order for an ordinary person to communicate with deaf

*Computer Science Department Mansoura University, Egypt
**Mansoura University, Egypt
***Computer Science Department Mansoura University, Egypt

people, a translator is usually needed to translate sign language into natural language [2, 3]. Furthermore, this translator should also be trained to translate speeches to sign Language.

Deaf community has been accustomed to conducting most of its daily affairs in isolation and only with people capable of understanding sign language. This isolation deprives this sizable segment of the society from proper socialization, education, and aspiration to career growth. This lack of communications hinders the deaf community from deploying their talents and skills in benefiting the society at large [4].

This paper addresses this problem by proposing a framework for a proposed pattern recognition system to support communication between the deaf community and rest of the society. The system receives the gesture video from the deaf person and converts it to the corresponding spoken word. It also receives the spoken word from the normal person and converts it to the corresponding gesture.

2. RELATED WORK

Attempts to automatically recognize sign language began to appear in the literature in the 90s. Research on hand gestures can be classified into two categories first category, relies on electromechanical devices that are used to measure the different gesture parameters such as the hand's position, angle, and the location of the fingertips. Systems that use such devices are usually called glove-based systems [5]. Major problems with such systems are force the singer to wear cumbersome and inconvenient devices. As a result, the way by which the user interacts with the system will be complicated and less natural.

The second category exploits machine vision and image processing techniques to create visual based hand gesture recognition systems. Visual-based gesture recognition systems are further divided into two categories. The first one relies on using specially designed gloves with visual markers called ‘‘visual-based gesture with glove–markers’’ that help in determining hand postures [6-8]. But using gloves and markers do not provide the naturalness required in human–computer interaction systems. Besides, if colored gloves are used, the processing complexity is increased.

Alternatively, the second kind of visual based hand gesture recognition systems can be called ‘‘pure visual based gesture’’ (i.e. visual-based gesture without glove–markers). This type tries to achieve the ultimate convenience naturalness by using images of bare hands to recognize gestures. Many feature extraction [9-11] methods and pattern recognition algorithms have widely been applied to characterize information on gesture images and videos for computer-aided sign language recognition [12-14].

Zahedi et al. [15] have presented an appearance-based sign language recognition system it uses a weighted combination of different geometric features (the hand area, the length of the hand border and the x and y coordinates of the center of gravity) and different appearance-based features (skin color intensity, and different kinds of first- and second-order derivatives) to recognize segmented American sign language, words.

Rousan et al. [16] have introduced a recognition system for Arabic sign language. The recognition system has been composed of several stages as follows: (1) video capturing, (2) segmentation, (3) Background removal, (4) Feature extraction using discrete cosine transform (DCT), and (5) Gesture recognition using hidden Markov models (HMMs).

Zaki and Shaheen [17] have presented a combination of vision appearance based features in order to enhance the recognition of underlying signs. Starting from the fact that a sign language is based on the four components (hand configuration, place of articulation, hand orientation, and movement). Three features were selected to be mapped to these four components. Two of these features are newly: kurtosis position and principal component analysis (PCA). The third feature was motion chain code that represented the hand movement.

Yun, Lifeng, and Shujun [18] have presented a hand gesture recognition method based on multi-feature fusion and template matching. The method detects hand-shaped contour region and obtains the maximum contour according to skin color feature, by extracting angle count, skin color angle, and non-skin color angle in combination with Hu invariant moments features of the largest hand shaped region for sample training. Euclidean distance template matching technique was applied for hand gesture classification and recognition.

On the other hand, during the last two decades, there have been important advances in the technological areas that support the implementation of an automatic speech to sign language translation systems.

San-Segundo et al. [19], have presented a system for Spanish to sign language translation system in a real domain. The system was made up of:

(1) speech recognizer for decoding the spoken utterance into a word sequence, (2) natural language translator for converting a word sequence into a sequence of signs belonging to the sign language, and (3) 3D avatar animation module for playing back the hand movements.

Foong, Low, and La [20] have presented a voice to sign language translation system for Malaysian deaf people. The main components are the sound recording component (with a supporting sound/voice training algorithm), digital signal processing component (with a supporting mel frequency cepstral coefficients (MFCC) algorithm counterparts), and the vector quantization component (supported by its matching sub component).

This paper motivates to serve both the deaf community and the normal society by fascinating communication between them. Therefore, it presents a recognition system for Arabic sign language based on the digital processing of both video and speech. The structure of the proposed system is presented in the following section.

3. METHODOLOGY

The proposed system includes two main modules: (1) video, and (2) sound processing. Each module comprises the following functions: (1) signal acquisition, (2) signal pre-processing; (3) signal analysis; and (4) training and testing. Fig. 1 shows the main block diagram of the proposed system.

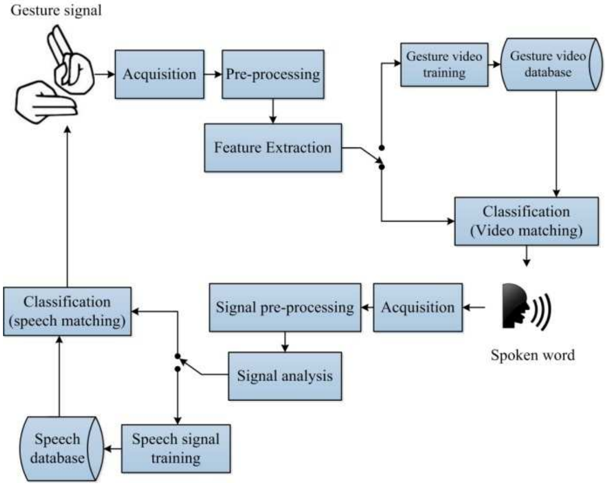

Fig 1: Block diagram of the proposed system

3.1. Video processing

3.1.1. Video acquisition

In the video capturing stage, a single digital camera was used to acquire the gestures from signers in a video format. At this stage the video is saved in the AVI format in order to be analyzed in latter stages. The signers will use bare hands, as the proposed system is machine vision based; i.e. no need for any gloves or any type of assisting devices.

3.1.2. Video pre-processing

At the pre-processing stage, the video is divided into frames then the following two steps are implemented:

a) Key frame extraction

In this section, the method of key frame extraction is discussed. This method is to extract key frames based on the change of shape information of each frame. The moments of inertia are used as a metric for providing the shape information. The basic idea is to compute the frame differences (moments of inertia differences) based on some criteria and then discard the frames whose difference with the adjacent frames are less than a certain threshold. The k frame do not become the new key frame until the distance between the $k-\text{th}$ and $(k-1)-\text{th}$ frame exceed a specific threshold.

Moments of inertia difference

Moments of inertia are frequently used in image processing as compact image descriptors that can differentiate an image from another image [21]. Hue, Saturation, Value (HSV) color space is used for moments of inertia difference computation which has the ability to provide intuitive representation of color closer to human perception. In this work, the three moments of inertia (mean, variance, skewness) are used to compute the 9 moments from each section of a frame (3 for each color channel). Again, a frame is divided into T_S number of sections of size $p\times q$ each. For a frame $F(t)$, the mean, variance, and skewness are computed for the three color channels of every section as [22]:

$$\overline{F(t)}_{s,c} = \frac{1}{p\,X\,q} \sum_{i=1}^{p} \sum_{j=1}^{q} F(t)_{i,j} \qquad (1)$$

$$\sigma^2 (F(t))_{s,c} = \frac{1}{p\,X\,q} \sum_{i=1}^{p} \sum_{j=1}^{q} \left(F(t)_{i,j} - \overline{F(t)}_{s,c} \right)^2 \qquad (2)$$

$$\gamma(F(t))_{s,c} = \frac{1}{p\,X\,q} \frac{\sum_{i=1}^{p} \sum_{j=1}^{q} \left(F(t)_{i,j} - \overline{F(t)}_{s,c} \right)^3}{\left(\sigma^2 (F(t))_{s,c} \right)^{3/2}} \qquad (3)$$

Where; $\overline{F(t)}_{s,c}$, $\sigma^2 (F(t))_{s,c}$ and $\gamma(F(t))_{s,c}$ are the mean, variance, and skewness values of color channel c in section s respectively. Finally, these values are combined to form a moments of inertia feature vector μ_t of frame $F(t)$. The size of the vector is $9 \times T_S$. The moments of inertia difference measure between two frames $F(t)$ and $F(t+1)$ is computed by using the Euclidean distance between the respective feature vectors.

$$\mu(F(t), F(t+1)) = \sqrt{\sum_{i=1}^{9T_S} \left(\mu_t(i) - \mu_{t+1}(i) \right)} \qquad (4)$$

b) Segmenting skin color from key frames

In sign languages, hand gestures represent the word meaning intended by the signer. Thus, by observing a sequence of hand gestures, extracting suitable features the required sign could be recognized. Therefore, a proposed method is presented to segment the face and the hand from the gesture image as follows.

Step1: Read the gesture image.

Step2: Convert image from RGB to the following two different color spaces, $YCbCr$ and HSV. The $YCbCr$ color space includes luminance component $(Cb \text{ and } Cr)$ and chromatic or pure color components (Y). This color space is separated from normalized RGB by the following transformation [23]:

$$\begin{bmatrix} Y \\ Cb \\ Cr \end{bmatrix} = \begin{bmatrix} 16 \\ 128 \\ 128 \end{bmatrix} + \begin{bmatrix} 65.481 & 128.553 & 24.966 \\ -37.797 & -74.203 & 112.00 \\ 112.00 & -93.786 & -18.214 \end{bmatrix} \begin{bmatrix} R \\ G \\ B \end{bmatrix} \quad (5)$$

The HSV color space includes hue (H), saturation (S) and value (V) used to describe color, in which humans experience and describe color sensations. The RGB color model can be converted to HSV model by the following equations [24]:

$$H = \begin{cases} 0 & \text{if } L_b = L_s, \\ \dfrac{\pi(G-B)}{3(L_b - L_s)} & \text{if } L_b = R \text{ and } G \geq B, \\ 2\pi + \dfrac{\pi(G-B)}{3(L_b - L_s)} & \text{if } L_b = R \text{ and } G < B, \\ 2\pi/3 + \dfrac{\pi(B-R)}{3(L_b - L_s)} & \text{if } L_b = G, \\ 4\pi/3 + \dfrac{\pi(R-G)}{3(L_b - L_s)} & \text{if } L_b = B, \end{cases} \quad (6)$$

$$S = \begin{cases} 0 & \text{if } L_b = 0, \\ \dfrac{L_b - L_s}{L_b} & \text{otherwise}, \end{cases} \quad (7)$$

$$V = L_b \quad (8)$$

Where

$L_b = max\left(R, G, B\right) and\ L_s = min\left(R, G, B\right)$ for each pixel.

Step3: According to the literature [25, 26], the segmentation of skin pixels can be performed by using the chromatic red and blue component values (Cr and Cb). In this paper, the following threshold which combines the $YCbCr$ and HSV color spaces, is used to accurately detect the skin color pixels and ignore any other pixels. For each detected skin pixel, the values fall within the ranges of

$$0.20 \leq S \leq 0.68 \& 0 \leq H \leq 0.2 \& 97.5 \leq Cb \leq 142.5 \& 143 \leq Cr \leq 176$$

Step4: Removing any remain small connected pixels using the morphological operator imopen.

Step5: Filling the holes within the skin. These holes are indeed the background pixels which are not connected to the image border. They are filled easily using a dilation of the skin (bright) pixels in the image resulted from the previous step. Obviously, the dilation is authorized only on the boundaries of holes. Fig.2 shows the result of skin color detection.

Step6: Face and hand extraction. The purpose of the segmentation in the proposed system is to extract the face and hand from the gesture image. After detecting skin components, a connected component labeling [27] is used where subsets of connected image components are uniquely labeled. This algorithm scans the image, labeling the underlying pixels according to a predefined connectivity scheme and the relative values of their neighbours. Only large components are taken into consideration. Three components represent the two hands and the face as shown in Fig. 3.

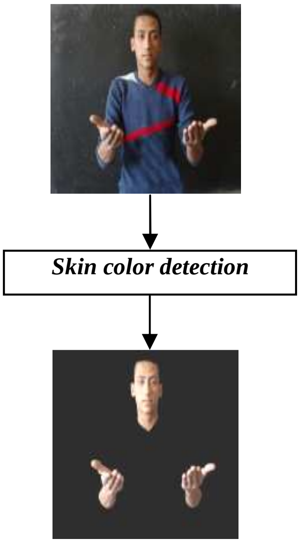

Fig 2: Result of skin color detection

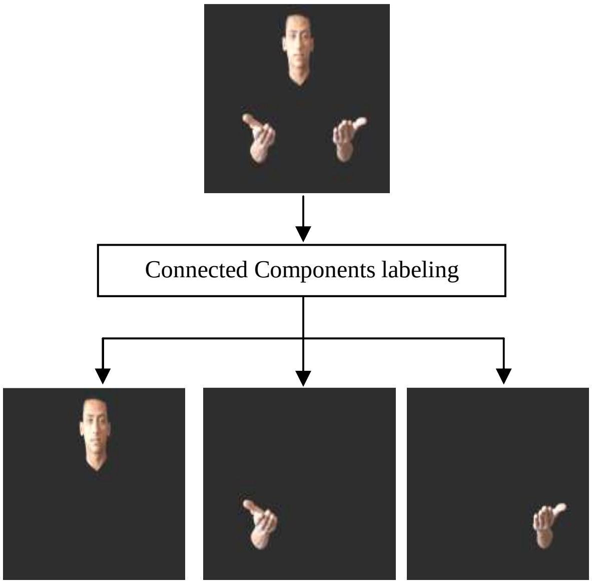

Fig 3: Result of connected components labeling

3.1.3. Video feature extraction

a) Two dimensional Wavelet packet decomposition

The wavelet transform (WT) was introduced at the beginning of the 1980s by Morlet, who used it to evaluate seismic data [28]. Wavelets provide an alternative to classical Fourier methods for one and multi-dimensional data analysis and synthesis, and have numerous applications both within mathematics (e.g., to partial differential operators) and in areas as diverse as physics, seismology, medical imaging, digital image processing, signal processing and computer graphics and video.

The main advantage of wavelets is that they have a varying window size, wide for slow frequencies and narrow for the fast ones, thus leading to an optimal time–frequency resolution in all frequency ranges. Furthermore, owing to the fact that windows are adapted to the transients of each scale, wavelets lack of the requirement of stationarity [29].

Wavelet decomposition uses the fact that it is possible to resolve high frequency components within a small time window, and only low frequencies components need large time windows. This is because a low frequency component completes a cycle in a large time interval whereas a high frequency component completes a cycle in a much shorter interval. Therefore, slow varying components can only be identified over long time intervals but fast varying components can be identified over short time intervals. Wavelet decomposition can be regarded as a continuous time wavelet decomposition sampled at different frequencies at every level or stage [30].

Two-dimensional (2D) (one along x-axis and the other along y-axis) discrete wavelet transform (DWT) can be implemented using digital filters and down samplers. In this way, we can apply convolution of low and high pass filters to the original data, and the image can be decomposed in specific sets of coefficients at each level of decomposition. Fig. 4 shows the 2D-DWT of image at ‘‘level 1” of decomposition. In the figure, the low pass filter is denoted by G_0 while the high pass filter is denoted by H_0. The image is first filtered along the x-direction, resulting in $f_l(x,y)$ and a high pass image $f_h(x,y)$. As the bandwidth of $f_l(x,y)$ and $f_h(x,y)$ is half along the ‘‘x” direction, each of the filtered images can be down sampled in ‘‘x” direction by 2 without loss of any information. The down sampling is accomplished by dropping every other filtered value. Both $f_l(x,y)$ and $f_h(x,y)$ are filtered along y-axis resulting in four sub-images [31, 32]. Again the sub-images are down sampled by 2 along the y-direction. According to the procedure, the image can be transformed into four sub-images, namely:

- f_{ll} sub-image: This is the trend image. Both horizontal and vertical directions (Approximation image).
- f_{lh} sub-image: This is partial detail image and the horizontal direction has low frequencies and the vertical one has high frequencies
- (Vertical detail image). f_{hl} sub-image: This is partial detail image and the horizontal direction has high frequencies and the vertical one has low frequencies (Horizontal detail image).
- f_{hh} sub-image: This is the detail image in both, horizontal and vertical directions (Diagonal detail image).

The approximations are the low-frequency components of the signal and details are high-frequency components.

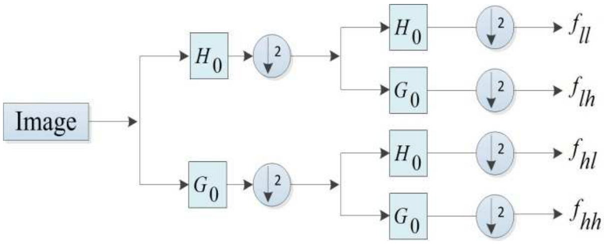

Fig 4: Two-dimensional discrete wavelet transform.

The wavelet packet method is an expansion of classical wavelet decomposition that presents more possibilities for signal processing [33]. The structure of wavelet packet transform (WPT) is similar to DWT. Both have the framework of multi-resolution analysis. In WT, signals split into a detail and an approximation. The approximation obtained from first-level is split into new detail and approximation and this process is repeated. Because of the fact that WT decomposes only the approximations of the signal, it may cause problems while applying WT to in certain applications where the important information is located in higher frequency components [34].

The main difference between WT and WPT is that WPT splits not only approximations but also details. The top level of the WPT is the time representation of the signal, whereas, the bottom level has better frequency resolution (Learned & Willsky, 1995). Thus, with the use of WPT, a better frequency resolution can be obtained for the decomposed signal. In addition, the use of WPT extracts much more features about the signal.

In 2D discrete wavelet packet transform, an image is decomposed into one approximation and three detail images. The approximation and the detail images are then decomposed into a second-level approximation and detail images, and the process is repeated. The standard 2D-DWPT can be implemented with a low-pass filter h and a high-pass filter g [35]. The 2D-DWPT of an $N \times M$ discrete image x up to level $P+1$ $(P \leq \min(\log_2 N, \log_2 M))$ is recursively defined in terms of the coefficients at level p as follows:

$$C^{p+1}_{4k,(i,j)} = \sum_{m}\sum_{n} h(m)h(n)C^{p}_{k,(m+2i,n+2j)}, \qquad (9)$$

$$C^{p+1}_{4k,(i,j)} = \sum_{m}\sum_{n} h(m)g(n)C^{p}_{k,(m+2i,n+2j)}, \qquad (10)$$

$$C^{p+1}_{4k+2,(i,j)} = \sum_{m}\sum_{n} g(m)h(n)C^{p}_{k,(m+2i,n+2j)}, \qquad (11)$$

$$C^{p+1}_{4k+3,(i,j)} = \sum_{m}\sum_{n} g(m)g(n)C^{p}_{k,(m+2i,n+2j)}, \qquad (12)$$

where $C^{0}_{0,(i,j)}$ is the image X . At each step, the image C^{p}_{k} is decomposed into four quarter-size images C^{p+1}_{4k}, C^{p+1}_{4k+1}, C^{p+1}_{4k+2}, C^{p+1}_{4k+3} A two-level wavelet packet decomposition is illustrated in Fig. 5.

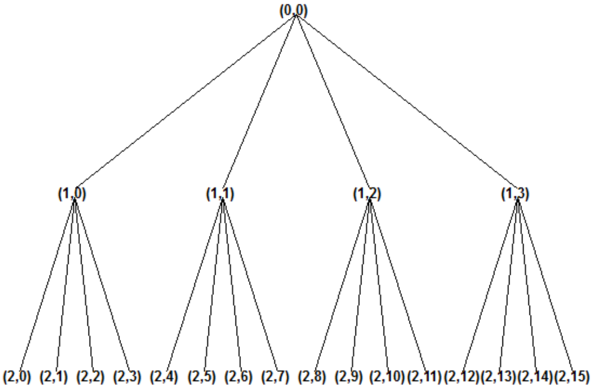

Fig 5: 2D wavelet packet tree for a two level decomposed image

The energy and entropy-based features corresponding to the terminal nodes of the decomposition tree of both the gesture image is calculated. The energy and entropy-based measures are calculated as follows.

Entropy feature. An Entropy-based criterion describes information-related properties for an accurate representation of a given signal. Entropy is a common concept in many fields, mainly in image processing and signal processing. The Shannon entropy is measured as follow [30],

$$E\left(f_{texture}\right) = \sum_{x}\sum_{y}\left|f\left(x,y\right)\right|^{p} \qquad (13)$$

Where $f(x,y)$ indicates the texture pixel at (x,y) position and p is the power and must be such that $1 \le p < 2$.

Energy feature: The energy distribution has important discriminatory properties for texture images and as such can be used as a feature for texture matching Energy of each channel can be computed as follow [36],

$$\sigma_p^2(k) = \sum_x \sum_y [C_k^p(x,y)]^2 \qquad (14)$$

Where $\sigma_p^2(k)$ is the energy of the texture projected to the subspace at node (p,k).

b) Shape features

Shape feature plays a vital role in object detection and recognition. Object shape features provide robust and efficient information of objects in order to identify and recognize them. Shape features are considered very important in describing and differentiating the objects in an image [37].

Hu invariant moments

Invariant moment theory is an important element in pattern recognition and computer vision. Common region-based invariant moment theory was first proposed by Hu [38], including the definition of continuous function moments and the basic nature of the moment, which was given a specific translation, rotation, scaling invariance of seven invariant moment expression. Since the proposal of Hu invariant moments, they have been applied to the image, character recognition and industrial quality control and many other fields. Here is the definition of Hu moments of digital image.

Let $f(x,y)$ is a digital image; the $(p+q)$ order moment is defined as [18]

$$m_{pq} = \sum_x \sum_y x^p y^q f(x,y) \; (p,q) = 0,1,2,..... \qquad (15)$$

The $(p+q)$order central moment is defined as

$$\mu_{pq} = \sum_x \sum_y (x-\bar{x})^p (y-\bar{y})^q f(x,y) \qquad (16)$$

In which

$$m_{01}=\sum_{x}\sum_{y} yf(x,y), m_{10}=\sum_{x}\sum_{y} xf(x,y) \qquad (17)$$

$$\bar{x}=\frac{m_{10}}{m_{00}}, \bar{y}=\frac{m_{01}}{m_{00}}, m_{00}=\sum_{x}\sum_{y} f(x,y) \qquad (18)$$

$(p+q)$order normalized central moments defined as

$$\eta_{pq}=\frac{\mu_{pq}}{\mu_{00}^{r}} \quad r=\frac{p+q}{2} \quad p+q=2,3,4,.... \qquad (19)$$

Seven invariant moments constituted by the linear combination of the second and third order central moments, the specific expression is as follow

$$\phi_1=\eta_{20}+\eta_{02}$$

$$\phi_2=(\eta_{20}-\eta_{02})^2+4\eta_{11}^2$$

$$\phi_3=(\eta_{30}-3\eta_{12})^2+\left(3\eta_{21}-\eta_{03}\right)^2$$

$$\phi_4=\left(\eta_{30}+\eta_{12}\right)^2+\left(\eta_{21}+\eta_{03}\right)^2 \qquad (20)$$

$$\phi_5=\left(\eta_{30}-3\eta_{12}\right)\left(\eta_{30}-\eta_{12}\right)\left[\left(\eta_{30}+\eta_{12}\right)^2-3\left(\eta_{21}+\eta_{03}\right)^2\right]$$
$$+\left(3\eta_{21}-\eta_{03}\right)\left(\eta_{21}+\eta_{03}\right)\left[3\left(\eta_{30}+\eta_{12}\right)^2-\left(\eta_{21}+\eta_{03}\right)^2\right]$$

$$\phi_6=\left(\eta_{20}-\eta_{02}\right)\left[\left(\eta_{30}+\eta_{12}\right)^2-\left(\eta_{21}+\eta_{03}\right)^2\right]+4\eta_{11}\left(\eta_{30}+\eta_{12}\right)\left(\eta_{21}+\eta_{03}\right)$$

$$\phi_7=\left(3\eta_{12}-\eta_{30}\right)\left(\eta_{30}+\eta_{12}\right)\left[\left(\eta_{30}+\eta_{12}\right)^2-3\left(\eta_{21}+\eta_{03}\right)^2\right]$$
$$+\left(3\eta_{21}-\eta_{03}\right)\left(\eta_{21}+\eta_{03}\right)\left[3\left(\eta_{03}+\eta_{12}\right)^2-\left(\eta_{12}+\eta_{03}\right)^2\right]$$

Moment features have the following physical meaning: the low-order moments describe the overall characteristics of the image, the zero-order moment reflects the target area, the first order moment reflects the target center of mass, second moment reflects the length of principal, auxiliary axis and the orientation angle of principal axis, higher moments describe the details of the image: as distortion and the kurtosis distribution of the target.

For each frame the 16 wavelet Entropy, the 16 wavelet Energy for textural feature and 7 moments for shape feature are respectively estimated from each frame .the constructed feature vector combines 39 components.

3.2. Speech processing

3.2.1 Signal acquisition

During the signal acquisition stage, speech signals were directly collected from normal people and saved as waveforms for subsequent analysis. A high quality microphone has been used to capture the speech signal according to the following parameter set (Sampling frequency 44100 Hz, resolution 16-bits/sample and recorded file format = *.wav).

3.2.2 Signal pre-processing

A number of audio cards add DC (Direct Current) components into the audio signal, Approaches used in digital signal processing are applied to compute some signatures. The DC component in the signal negatively affects the computation and may cause disturbance.

A pre-emphasis filter is typically a simple first order high pass filter. The purpose of pre-emphasis is to even the spectral energy envelope by amplifying the importance of high-frequency components and removing the DC component in the signal. The z-transform of the filter is shown in Eq.

$$H(z) = 1 - \alpha * z^{-1} \ , 0.9 \leq \alpha \leq 1.0 \qquad (21)$$

In the time-domain, the relationship between the output s'_n and the input s_n of the pre-emphasis .For fixed-point implementations a value of $\alpha = 15 / 16 = 0.9375$ is often used [39].

The pre-emphasized signal was divided into short frame segment using Hamming window. Fig. 6 represents the original speech signal. Fig.7 displays the signal after applying the pre-emphasizing filtering technique.

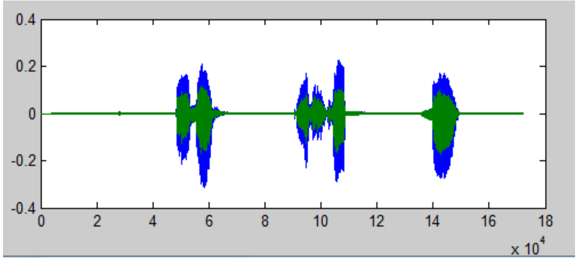

Fig 6: Sample form of recorded speech signal.

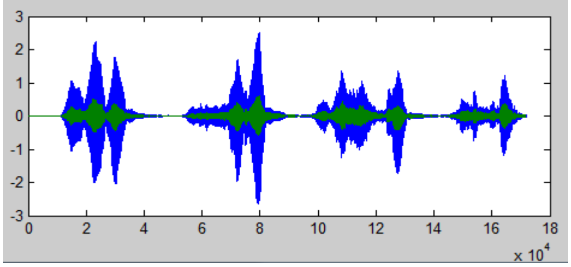

Fig 7: Speech after pre-emphasizing filtering technique.

3.2.3 Signal analysis

In all recognition systems, signal processing is carried out to convert the raw signals to some type of parametric representation. This parametric representation is then used for further analysis and processing and is called a feature. In this paper, the speech signals are represented by a combination of the mel-frequency cepstral coefficient (MFCC), and the timpral textural features.

Mel-frequency cepstral analysis

The feature extraction process of MFCC [40-42] includes:

Step1: The speech signals are blocked into short frames of N samples, with a predefined overlapping value (50% overlapping).

Step2: Each individual frame of a signal x is windowed so as to minimize the signal discontinuities at the beginning and at the end of each frame and thus the spectral distortion is minimized. The window is defined as given below:

$$w(n); \quad where \quad 0 \le n \le (N-1). \qquad (22)$$

N is the number of samples in each frame. The result of windowing is the signal $y(n)$ and is defined as,

$$y(n) = x(n)w(n), \quad where \quad 0 \le n \le (N-1), \qquad (23)$$

the hamming window $y(n)$, used in this work as shown in Fig. 8 is given by,

$$w(n) = 0.54 - 0.46 \cos\left[2\pi n / (N-1)\right], \; 0 \le n \le (N-1), \quad (24)$$

the purpose of the window is to favour samples towards the centre of the window. This characteristic coupled with the overlapping attempts to smoothen the varying parameters. Fig. 9 shows the result of this step.

Step3: Fast Fourier Transform (FFT) is applied to the windowed samples, which converts each frame of N samples from the time domain into the frequency domain, The FFT is defined on the set of N samples $\{x_n\}$ as:

$$X_n = \sum_{k=0}^{N-1} x_k e^{-2\pi jkn/N}, \; where \;\; n = 0, 1, 2 \; . \; . \; . \; N-1, \quad (25)$$

where x_n s are the complex numbers. The resulting sequence of x_n s is interpreted as given: (1) when $(n=0)$, it corresponds to zero frequency; (2) When $1 \le n \le (N/2-1)$, it corresponds to positive frequencies $(0 < f < F_S/2)$; and (3) When $N/2+1 \le n \le N-1$, it corresponds to negative frequencies $(-F_S/2 < f < 0)$. Here, F_S denotes the sampling frequencies. The obtained result is often referred to as ‘spectrum’ or ‘periodogram’.

Step4: The Mel-frequency warping is implemented. Psychophysical studies have shown that human perception of the frequency contents of sounds does not follow a linear scale. MFCC are based on the known variation of the human ear’s critical bandwidths with frequency. Thus the Mel-frequency scale is used which is the linear frequency spacing below 1000 Hz and a logarithmic spacing above 1000 Hz. The following approximate empirical relationship to compute the Mel frequencies (also called Mel’s) for a given frequency f expressed in Hz is as given below:

$$Mel(f) = 2595 \times \log(1 + f/700) \qquad (26)$$

in order to simulate the frequency warping process, we use a filter bank, one filter for each desired Mel-frequency component. That filter bank as shown in Fig. 10 has a triangular band-pass frequency response, and the spacing as well as the bandwidth is determined by a constant Mel-frequency interval.

Step5: The log Mel-frequency spectrum is converted back to time domain using Discrete Cosine Transform (DCT). The resultant is called the MFCC as shown in Fig. 11 and are calculated using:

$$c_n = \sum_{k=1}^{K} (\log S_k) \cos\left[n\left(k - \frac{1}{2}\right)\frac{\pi}{k} \right] \qquad (27)$$

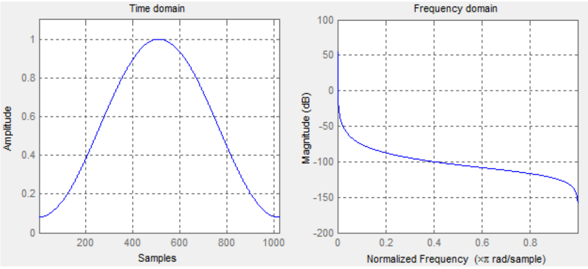

Fig 8: The time and frequency domain plots of a hamming window of a length is equivalent to 1024 bins applied in this paper.

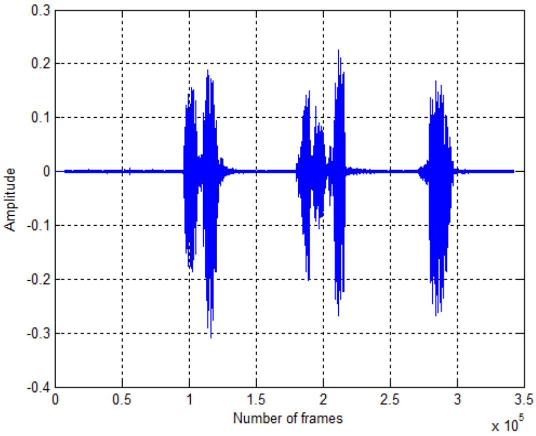

Fig 9: The windowing step of the speech signal.

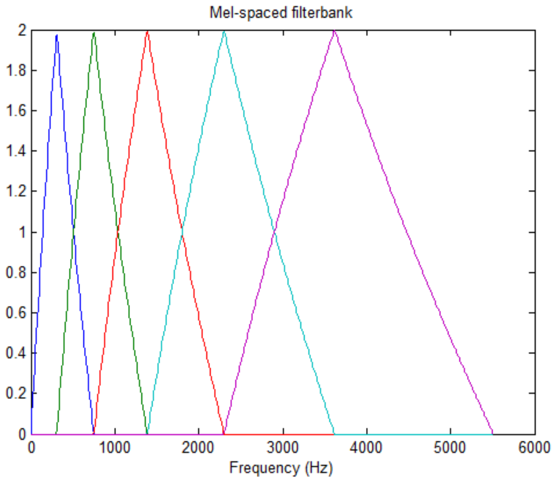

Fig 10: The Mel-spaced filter bank with 5 filters applied to the speech signals.

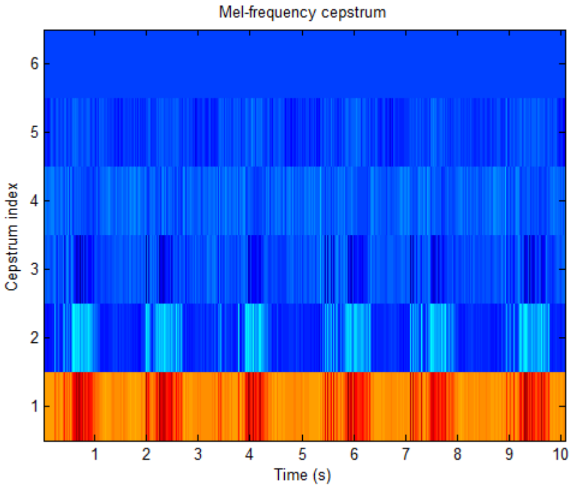

Fig 11: The MFCC of speech signal.

Timbral texture analysis

These features are based on the short time Fourier transform (STFT) and are calculated for every short-time frame of speech. The following low-level signal features, representing timbral texture, are used in this work:

Spectral centroid: Is the center of gravity of the magnitude spectrum of the STFT:

$$C_t = \frac{\sum_{n=1}^{N} M_t[n] * n}{\sum_{n=1}^{N} M_t[n]} \qquad (28)$$

where $M_t[n]$ is the magnitude of the Fourier transform at frame t and frequency bin n. The centroid is a measure of spectral shape and higher centroid values correspond to 'brighter' textures with more high frequencies. Fig.12 shows the spectral centroid of a speech signal.

Spectral rolloff: Is the frequency R_t below which 85% of the magnitude distribution is concentrated:

$$\sum_{n=1}^{R_t} M_t[n] = 0.85 * \sum_{n=1}^{N} M_t[n] \qquad (29)$$

The rolloff is another measure of spectral shape and yields higher values for high frequencies. Fig.13 shows the Spectral rolloff a speech signal.

Spectral flux: Is the squared difference between the normalized magnitudes of successive spectral distributions:

$$F_t = \sum_{n=1}^{N} (N_t[n] - N_{t-1}[n])^2 \qquad (30)$$

where $N_t[n]$ and $N_{t-1}[n]$ are the normalized magnitude of the Fourier transform at the current frame t, and the previous frame $t-1$, respectively. The spectral flux is a measure of the amount of local spectral change. Fig.14 shows the Spectral flux a speech signal.

Zero crossings rate: Is the rate of sign-changes of a signal, i.e., the number of times the signal changes from positive to negative or back, per time unit. It is defined as presented in the following equation.

$$Z_t = \frac{1}{2} \sum_{n=1}^{N} \left| sign(x[n]) - sign(x[n-1]) \right| \qquad (31)$$

where the sign function is 1 for positive arguments and 0 for negative arguments and $x[n]$ is the time domain signal for frame t. Time domain zero crossings provide a measure of the noisiness of the signal. Fig.15 shows the Zero crossings rate a speech signal.

For the feature extraction, the speech signal is split into a sequence of short-term frames of 23.22 ms (1024 samples per frame) with 50% overlapping. Each individual frame is windowed with a hamming analysis window. The first five MFCC, the four timbral textural featuers, are respectively estimated from each enframed signal, and two acoustic matrices are constructed. Therefore, for each speech signal, two feature sequences of length w are calculated. In order to extract semantic content information it is necessary to follow how those sequences change from frame to frame. To quantify this variation, a number of statistics (mean, standard deviation and skewness) have been calculated. Therefore, the constructed feature vector combines 27 components.

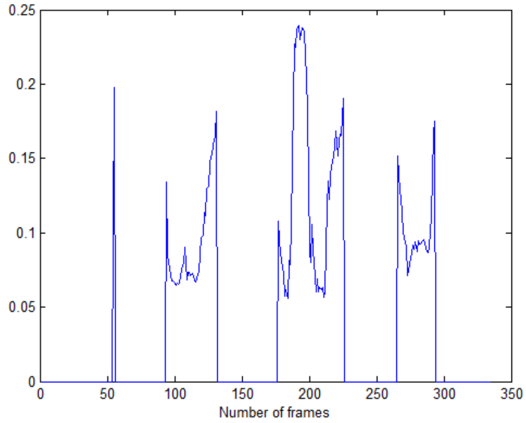

Fig 12: The spectral centroid of speech signal.

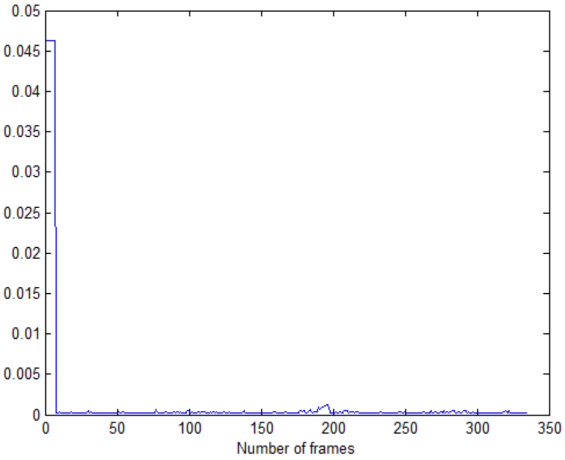

Fig 13: the Spectral rolloff of a speech signal.

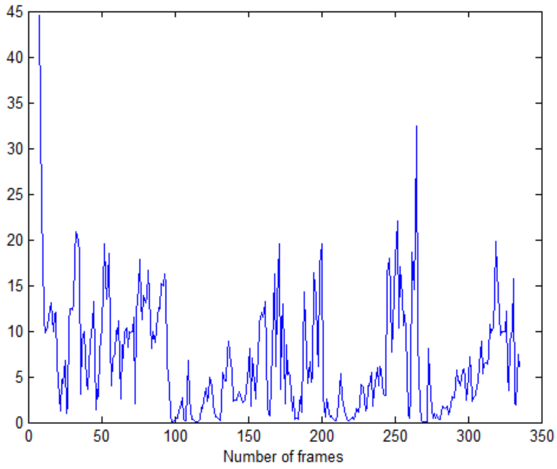

Fig 14: the Spectral flux of a speech signal.

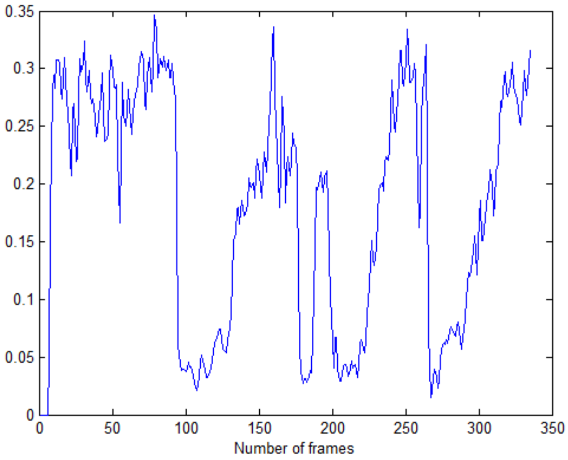

Fig 15: the Zero crossings rate of a speech signal.

4. Pattern recognition

In this stage, the proposed pattern recognition system is trained and tested using the gestures videos and their corresponding spoken words. Because there has been no serious attention to Arabic sign language recognition, there are no common databases available for researchers in this field. Therefore, a gesture database image has been built with reasonable size.

The data set used for training and testing the pattern recognition system consists of color images for all of the 25 signs for 25 word used in the experiment, these 25 signs are shown in Fig.16. Also 15 samples for each sign were taken from 15 different volunteers. The samples were taken from different distances by digital camera, and with different orientations. Moreover, 15 recorded spoken words corresponding to each sign were taken from 15 different normal persons to build the system database. For each sign or spoken word, 10 out of 15 samples were used for training purpose, while the remaining five signs were used for testing. In this paper WED is used for training and testing the pattern recognition system.

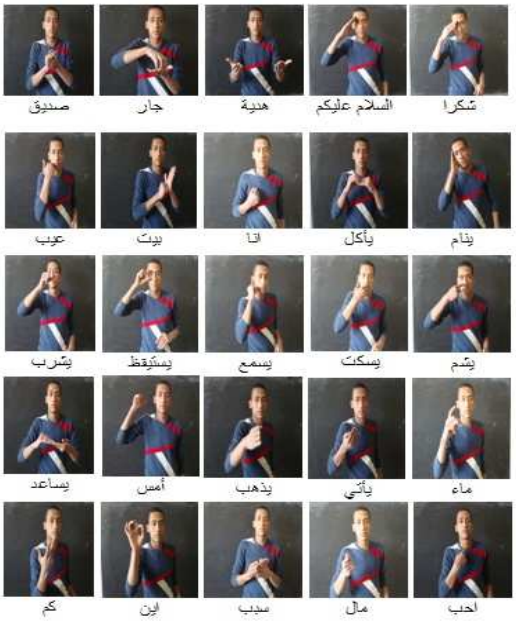

Fig 16: Arabic sign language words.

Weighted Euclidean distance

Many pattern recognition methods have been applied to the gesture classification problem. Generally speaking, linear classifiers require only simple arithmetic operations and have steady classification performance whereas non-linear classifiers, such as neural networks and SVMs, provide higher classification accuracy but involve much more complex calculation operations [43, 44]. In order to reach a high classification speed with a satisfying classification rate, a Weighted Euclidean Distance Classifier was adopted in this research work.

If Q is the feature vector of testing signal and I is the feature vector of the database, weighted Euclidean distance is computed as follows:

$$D(I,Q) = \sqrt{\sum_{i=1}^{n} wi\,(fiI - fiQ)^2} \qquad (32)$$

Where

n is the dimension of image feature. fiI , fiQ are i^{th} feature component of I and of Q respectively and wi is weight factor.

To calculate weights, is given by

$$w_i = \frac{N}{\sum_{K=1}^{N} \left(I_i^k - \overline{Q_i}\right)^2} \qquad (33)$$

Where

$$\overline{Q_i} = \frac{\sum_{k=1}^{N} I_i^k}{N} \qquad (34)$$

Where w_i denotes to the weight add to the component Q to balance the variations in the dynamic range, the value of k for which the function is minimum is selected as the matched image index. the value of n denotes to the dimension of the feature vector and the value of N denotes to the number of image in database .

5. Experimental results

A computer-based sound recognition system has been developed to support communication between the deaf community and the normal. The system can acquire, save, analyze, and both recognize the gesture videos of deaf person as well as the speech signals of the normal one. Firstly, it receives the gesture video from the deaf person and converts it to the corresponding spoken word. Secondly, it receives the spoken word from the normal person and converts it to the corresponding gesture. The graphical user interface (GUI) of the system is shown in Fig.19.

From the figure the query gesture video or spoken word can be selected, analyzed, and tested. The button titled 'Load video' reads the gesture video. The button titled 'Video to frames' divides video to frames. The button titled 'Key frame' extracts key frames from divided frames. The button titled 'Testing' tests the pattern and converts it to spoken words. The button titled 'training' update gesture video system database. The button titled Load sound' reads spoken word. The button titled sound pre-processing' remove the noise from the sound signal. The button titled 'frame-blocking' divides sound to frames. The button titled 'testing' tests the pattern and converts it to gesture video. The button titled 'training' update spoken words system database.

In order to assess the generalization capabilities of the system, one the most common evaluation measures in content-based retrieval is used. This performance measure is called, precision and recall [45], usually presented as a precision vs. recall graph (P-R graph). Thus, to give a just and impersonal evaluation of the retrieval efficiency of the proposed feature extraction method, the P-R graph is used as a criterion of evaluating performance:

$$precision = \frac{Number\ of\ relevant\ images\ retrieved}{Total\ number\ images\ retrieved} \qquad (35)$$

$$recall = \frac{Number\ of\ relevant\ images\ retrieved}{Total\ number\ of\ relevant\ images\ in\ the\ database} \qquad (36)$$

In the simulation, the system is respectively tested for the two types of features presented in this paper: (1) the combined DWP and invariant moments features extracted from the gesture videos, and (2) the combined MFCC and timbral textural features extracted from spoken words.

Fig.17, demonstrates the retrieval performance obtained by the speech features. From the figure, we can see that the retrieval performance obtained using the MFCC combined to the timbral texture features outperforms those obtained using the single methods.

Fig.18, demonstrates the retrieval performance obtained by the video gestures features. From the figure, we can see that the retrieval performance obtained using the two kinds of features together outperforms those obtained using the texture or shape features singly because only using texture or shape information can't adequately describe the video.

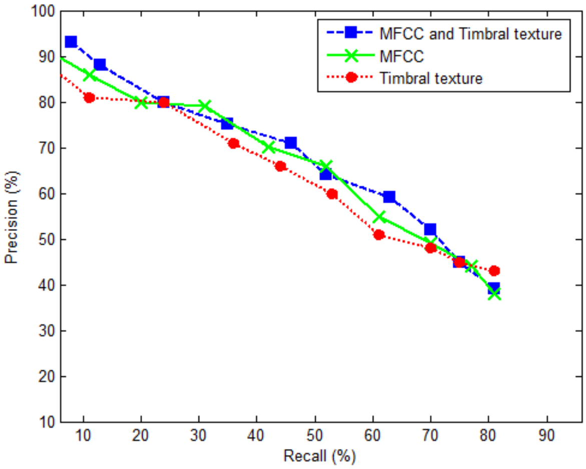

Fig 17: Retrieval performance comparisons for the three video feature extraction methods.

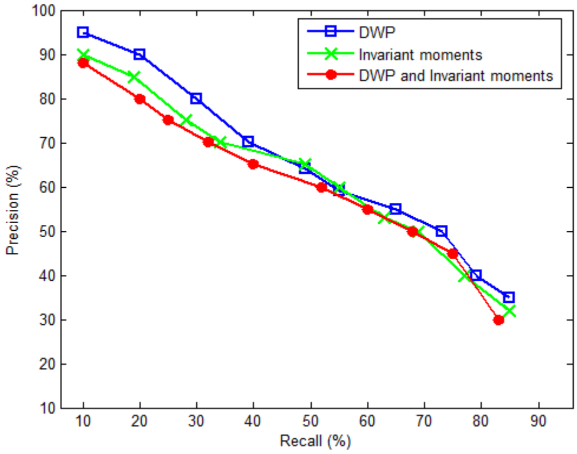

Fig 18: Retrieval performance comparisons for the three speech signal analysis methods.

6. CONCLUSIONS

The sign language is the fundamental communication method between people who suffer from hearing defects. In this paper, a computer-based sound recognition system has been developed to support communication between the deaf community and the normal. The system can acquire, save, analyze, and both recognize the gesture videos of deaf person as well as the speech signals of the normal one. The proposed system includes two main modules: (1) video, and (2) sound processing. Each module comprises the following functions: (1) signal acquisition, (2) signal pre-processing; (3) signal analysis; and (4) training and testing. Firstly, the system receives the gesture video from the deaf person and converts it to the corresponding spoken word. Secondly, it receives the spoken word from the normal person and converts it to the corresponding gesture. Results show that the designed system is effective in supporting communication between the deaf community and the normal persons.

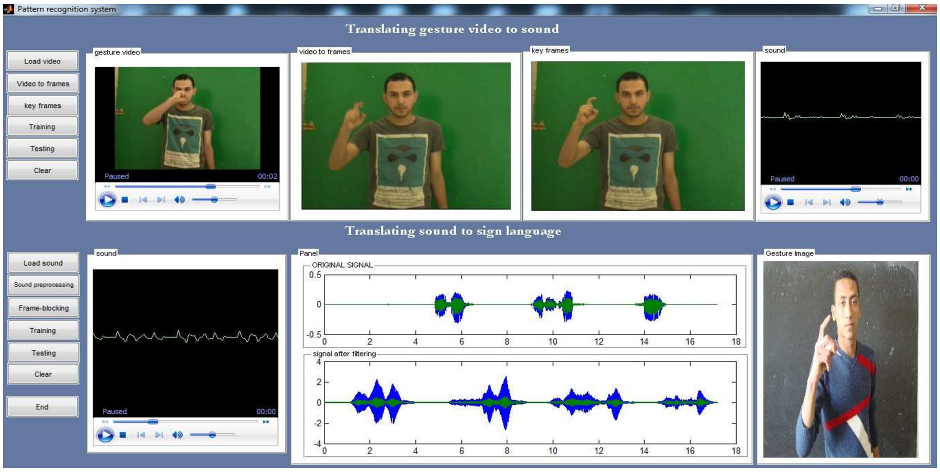

Fig 19: The graphical user interface (GUI) of the proposed system.

REFERENCES

1. Al-Jarrah,O., and Halawani,A. 2001. Recognition of gestures in Arabic sign language using neuro-fuzzy systems. Artificial Intelligence, 133, 117–138.
2. International Bibliography of Sign Language. 2005. http://www.signlang.uni-hamburg.de/bibweb/F-Journals.html.
3. International Journal of Language & Communication Disorders. 2005. Available from http://www.newcastle.edu.au/renwick/ROL/Jnlcontents/000mgmgf.htm.
4. Shanableh, T., and Assaleh, K. 2011. User-independent recognition of Arabic sign language for facilitating communication with the deaf community. Digital Signal Processing, 21, 535–542.
5. Grimes, G. 1983. Digital data entry glove interface device, Patent 4,414,537, AT & T Bell Labs.

6. Fels, S., and Hinton, G. 1993. GloveTalk: a neural network interface between a DataGlove and a speech synthesizer. IEEE Transactions on Neural Networks, 4, 2–8.
7. Dorner, B., and Hagen, E. 1994. Towards an American sign language interface. Artificial Intelligence Review, 8(2–3), 235–253.
8. Starner, T. 1995. Visual recognition of American sign language using hidden Markov models. Master’s thesis, Massachusetts Institute of Technology.
9. Teng, X., Wu, B., Yu, W., and Liu, C. 2005. A hand gesture recognition system based on local linear embedding. Journal of Visual Languages and Computing,16, 442–454.
10. Kelly, D., McDonald, J., and Markham, C. 2010. A person independent system for recognition of hand postures used in sign language. Pattern Recognition Letters, 31, 1359–1368.
11. Shen, X., Hua, G., Williams, L., and Wu, Y. 2012. Dynamic hand gesture recognition: An exemplar-based approach from motion divergence fields. Image and Vision Computing, 30, 227–235.
12. Lee, Y., and Tsai, C. 2009. Taiwan sign language (TSL) recognition based on 3D data and neural networks. Expert Systems with Applications, 36, 1123–1128.
13. Krishnaveni, M., and Radha, V. 2012. Classifier fusion based on Bayes aggregation method for Indian sign language datasets. Procedia Engineering, 30, 1110 – 1118.
14. Stergiopoulou, E., and Papamarkos, N. 2009. Hand gesture recognition using a neural network shape fitting technique. Engineering Applications of Artificial Intelligence, 22, 1141–1158.
15. Zahedi, M., Dreuw, P., Rybach, D., Deselaers, T., Ney, H., 2006. Using geometric features to improve continuous appearance-based sign language recognition. British Machine Vision Conference (BMVC), 3, 1019-1028.
16. AL-Rousan, M., Assaleh, K., and Tala’a, A. 2009. Video-based signer-independent Arabic sign language recognition using hidden Markov models. Applied Soft Computing, 9, 990–999.
17. Zaki, M., and Shaheen, S. 2011. Sign language recognition using a combination of new vision based features. Pattern Recognition Letters, 32, 572–577.
18. Yun, L., Lifeng, Z., and Shujun, Z. 2012. A Hand Gesture Recognition Method Based on Multi-Feature Fusion and Template Matching. Procedia Engineering, 29, 1678-1684.

19. San-Segundo, R., Barra, R., Co´rdoba, R., D’Haro, L., Ferna´ndez, F., and Ferreiros, J. 2008. Speech to sign language translation system for Spanish. Speech Communication, 50, 1009–1020.
20. Foong, O., Low, T., and La, W. 2009. V2S: Voice to Sign Language Translation System for Malaysian Deaf People. Lecture Notes in Computer Science, 5857, 868-876.
21. Flusser, J., Zitova, B., and Suk, T. 2009. Moments and Moment Invariants in Pattern Recognition. Wiley & Sons Ltd, 312.
22. Ejaz, N., Tariq, T., and Baik, S. 2012. Adaptive key frame extraction for video summarization using an aggregation mechanism. Visual Communication and Image Representation, 23, 1031–1040.
23. Aibinu, A., Shafie, A., and Salami, M. 2012. Performance Analysis of ANN based YCbCr Skin Detection Algorithm. Procedia Engineering, 41, 1183 – 1189.
24. Chen, Y., Hu, K., and Ruan, S. 2012. Statistical skin color detection method without color transformation for real-time surveillance systems. Engineering Applications of Artificial Intelligence, 25, 1331–1337.
25. Hiremath, P., and Danti, A. 2006. Detection of multiple faces in an image using skin color information and lines-of-separability face model. International Journal of Pattern Recognition and Artificial Intelligence, 20(1), 39–61.
26. Wang, Y., Yuan, B. 2001. A novel approach for human face detection from color images under complex background. Pattern Recognition, 34, 1983-1992.
27. Kang, S., Nam, M., and Rhee, B. 2008. Color Based Hand and Finger Detection Technology for User Interaction. International Conference on Convergence and Hybrid Information Technology, 229 – 236.
28. Mertins, A. (1999). Signal analysis wavelets, filter banks, time-frequency transforms and applications, Wollongong, Australia.
29. Ha, Q., Tran, T., and Dissanayake, G. 2005. A wavelet- and neural network-based voice interface system for wheelchair control. International Journal of Intelligent Systems Technologies and Applications, 12(4), 49–65.
30. Sengur, A., Turkoglu, I., and Ince, M. 2007. Wavelet packet neural networks for texture classification. Expert Systems with Applications, 32, 527–533.
31. Kumar, S., and Kumar, D., k. (2005). Visual hand gestures classification using wavelet transform and moment based features. International Journal of Wavelets, Multiresolution and Information Processing, 3, 79–101.
32. Gonzalez, R. C., & Woods, R. E. (2002). Digital image processing. New Jersey: Prentice Hall.

33. Ekici, S., Yildirim, S., and Poyraz, M. 2008. Energy and entropy-based feature extraction for locating fault on transmission lines by using neural network and wavelet packet decomposition. Expert Systems with Applications, 34, 2937–2944.
34. Shinde, A. D. (2004). A wavelet packet based sifting process and its application for structural health monitoring, Master Thesis, Faculty of Worcester Polytechnic Institute, pp. 22–32.
35. Mallat, S., A. Wavelet Tour of Signal Processing, Academic Press, New York, 1999.
36. Huang, K., and Aviyente, S. 2006. Information-theoretic wavelet packet subband selection for texture classification. Signal Processing, 86, 1410–1420.
37. Iqbal, K., Odetayo, M., and James, A.2012. Content-based image retrieval approach for biometric security using colour, texture and shape features controlled by fuzzy heuristics. Journal of Computer and System Sciences,78, 1258–1277.
38. Hu, M., K． 1962. Visual Pattern Recognition by Moment Invariants． IRE Trans． Information theory , 8, 179–187．
39. Ai, O., C., Hariharan, M., Yaacob, S., and Chee, L., S. 2012. Classification of speech dysfluencies with MFCC and LPCC features. Expert Systems with Applications, 39, 2157–2165.
40. Bahoura, M. 2009. Pattern recognition methods applied to respiratory sounds classification to normal and wheeze classes. Computers in Biology and Medicine, 39 (9). 824–843.
41. Bahoura, M., and Pelletier, C. 2004. Respiratory sounds classification using cepstral analysis and Gaussian mixture models. In 26th Annual Conference of the IEEE EMBS, San Francisco, CA, September 1–5, pp. 9–12.
42. Calvo de Lara, J.R. 2005. A method of automatic speaker recognition using cepstral features and vectorial quantization. Springer-Verlag Berlin Heidelberg, 3773, 146-153.
43. Bian, Z., Q., and Zhang, G., X. 2002. Pattern Recognition (2nd Edition). Tsinghua University Press, Beijing .
44. Zhao, Z., Chen, X., Zhang, X., Yang, J., Tu, Y., Lantz, V., and Wang, K. 2007. Study on Online Gesture sEMG Recognition. Lecture Notes in Computer Science, 4681, 1257–1265.
45. Mueller, H., Mueller, W., Squire, D., Maillet, S.M., and Pun, T. 2001. Performance evaluation in content-based image retrieval: overview and proposals. Pattern Recognition Letters, 22(5), 593-601.