Applied Mathematics & Information Sciences
An International Journal

http://dx.doi.org/10.12785/amis/090615

Multimodal Arabic Speech Recognition for Human-Robot Interaction Applications

Alaa Sagheer[1,2,*]

[1] College of Computer Science and Information Technology, King Faisal University, Al Hofuf, Kingdom of Saudi Arabia
[2] Center for Artificial Intelligence and RObotics (CAIRO), Faculty of Science, Aswan University, Aswan, Egypt

Received: 15 Feb. 2015, Revised: 15 Apr. 2015, Accepted: 16 Apr. 2015
Published online: 1 Nov. 2015

Abstract: By the earliest motivation of building humanoid robot to take care of human being in the daily life, the researches of robotics have been developed several systems over the recent decades. One of the challenges faces humanoid robots is its capability to achieve audio-visual speech communication with people, which is known as human-robot interaction (HRI). In this paper, we propose a novel multimodal speech recognition system can be used independently or to be combined with any humanoid robot. The system is multimodal since it includes audio speech module, visual speech module, face and mouth detection and user identification all in one framework runs on real time. In this framework, we use the Self Organizing Map (SOM) in feature extraction tasks and both the k-Nearest Neighbor and the Hidden Markov Model in feature recognition tasks. Results from experiments are undertaken on a novel Arabic database, developed by the author, includes 36 isolated words and 13 casual phrases gathered by 50 Arabic subjects. The experimental results show how the acoustic cue and the visual cue enhance each other to yield an effective audio-visual speech recognition (AVSR) system. The proposed AVSR system is simple, promising and effectively comparable with other reported systems.

Keywords: Human-robot interaction, Multimodal interaction processing, Audio-visual speech recognition, Face detection, User identification, Lip reading

1 Introduction

Due to the increasing demands for the symbiosis between human and robots, humanoid robots (HRs) are expected to offer the perceptual capabilities as analogous as human being. One challenge of HRs is its capability of communication with people. Intuitively, (1) hearing capabilities (audio speech), (2) visual information (visual speech) and (3) user identity (UID) are essential to be combined with an HR. A system combines these modules in one framework and recovers the traditional defects of each module is urgently needed [1]. Recently, the author presented a novel system combines the visual speech recognition (VSR) module with the UID module in one framework [2]. In this paper, we add a third module of the audio speech recognition (ASR) to the system presented in [2] to yield an overall audio-visual speech recognition (AVSR) system. Needless to say that audio speech is very important medium for communication; therefore, a considerable part of current HRs applications is based on audio. Unfortunately, most of the current ASR technologies possess a general weakness; they are vulnerable to cope with the audio corruption. Thus, their performance would dramatically degrade under the real-world noisy environments. This undesired outcome stimulates a flourish development of AVSR systems [3]. In the same time, most of the current AVSR systems are neglecting the UID, which can be useful for further interaction and secure communication between the human and the robot. Accordingly, the proposed AVSR system integrates face detection, mouth detection, user identification and word recognition. It adapts the subject facial movements that cannot be avoided in real life. The systems scenario runs as follows: The subject sits in front of the computer, then the system starts to detect the subjects face, tracks the subjects face, detects the subjects mouth and identifies the subject. Once the subject utters a word, the system recognizes this word. Indeed, the addition of the acoustic element to the system presented in [2] represents an important upgrade for the visual element. Another difference between this paper and [2] is that here, we increase the number of words from 9 words

[*] Corresponding author e-mail: asagheer@aswu.edu.eg

in [2] to be 26 words. Of course, using large number of words makes our system more robust. The important contribution of the proposed system here is that it runs on real time. HRI applications may require such real time systems to enable interaction with the robot. A lot of applications can utilize this system, such as voice dialing, speech-to-text processing, health care systems for elderly and disabled, multi-media phones for hearing impaired, mobile phone interface for public spaces, recovery of speech from deteriorated or mute movie clips and security by video surveillance etc. The paper is organized as follows: Section 2 presents an overview of previous works. A motivation and outline of our work are provided in section 3. Section 4 concludes the previous VSR system presented in [2]. The enhancements of the VSR system [2] are given in section 5. The overall AVSR system is provided in section 6. Database is described in section 7. Experimental results and comparisons are provided in section 8. Section 9 concludes this paper and shows our future work.

2 Previous works

Although significant advances have been made in ASR technology, it is still a difficult problem to design an effective AVSR system can generalize well without 1- loss of features and 2- image restrictions, see [1,3,4]. In our thinking, the difficulty of AVSR research is due to the large appearance variability during lip movements and large hearing variability during the pronunciation of one word by the same person. In addition, appearance differences across subjects, differences in lip sizes and face features and differences in illumination conditions cause extra difficulty. This section gives a survey about traditional AVSR systems and relation with humanoid robot.

Traditional AVSR systems

According to the best of our knowledge, there are three generations of AVSR systems have been developed so far, as follows:

1. ”Offline” AVSR systems
   In the offline AVSR systems, the designer gathers a large number of data samples from different subjects through different and separated sessions. A part of this data is used to train the system and the other part is used to test it. Both training and testing phases are running offline and separately. The experiments of these earlier systems focus only visually on the users mouth region and neglect the rest of the face and acoustically on the phonemes of the uttered words, as it depicted in Figure 1.

   Developing of offline systems is started at 1990s till 2004. One of the pioneers of such systems is Luettin who used hidden Markov model (HMM) based active shape model (ASM) to extract active speech features set that includes derivative information and compared its performance with that of a static feature set [5,6]. Matthews et al. compared three image transform based methods (discrete cosine transform DCT, wavelet transform WT and principal component analysis PCA) with active appearance model (AAM) to extract features from lip image sequences for recognition using HMM [7,8]. Heckmann investigated different tactics to choose coefficients of DCT to enhance feature extraction [9]. Using asymmetrically boosted HMM, Yin et al. developed an automatic visual speech feature extraction to deal with their own ill-posed multi-class sample distribution problem [10]. Guitarte et al. compared between ASM and DCT for feature extraction task in an embedded implementation [11]. Hazen investigated several visual model structures, each of which provides a different means for defining the units of the visual classifier and the synchrony constraints between the audio and the visual stream [12]. The author of this paper presented an appearance based visual speech recognition system combines his approach Hyper-Column Model (HCM) as a feature extractor with HMM as a feature recognizer [13]. Sagheer systems performance evaluated using multiple sentences from two databases for Japanese [14] and Arabic [15] languages. Sagheer demonstrated that his approach outperforms DCT-base approach [13].

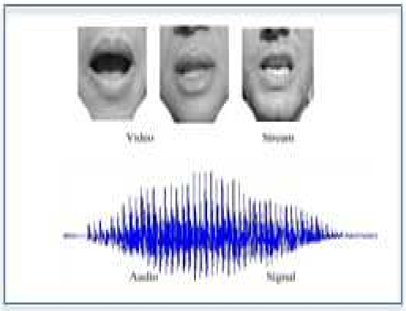


Fig. 1: Asynchronous between audio signal and visual frames.

2. ”Offline” multi-elements AVSR systems

   Later on, i.e. in the period 2004-2011, many visual elements have been included to the AVSR systems, however, they were still working offline. These elements are face detection, mouth detection, face recognition, gesture recognition and so on. For examples, Sanderson et al. evaluated several recent nonadaptive and adaptive techniques for reaching the verification of the combination of speech and face information in noisy conditions to perform speech recognition and face identification system [16]. Cetingl et al. presented a new multimodal speaker recognition system that integrates speech, lip texture and lip motion based on a combination of a well-known mel-frequency cepstral coefficients (MFCC) and 2D-DCT. The fusion of speech, lip texture and lip motion modalities is performed by a reliability weighted summation decision rule in [17]. Saitoh et al. presented an analysis of efficient lip reading methods for various languages. First, they applied active appearance model (AAM), and simultaneously extracted the external and internal lip contour. Then, the tooth and intraoral regions were detected. Various features from five regions were fed to the recognition process. They took four languages for the recognition target, and recorded 20 words per each language [18]. Deypir et al. presented a new method of boosting algorithm based on Multi-linear-Discriminant Analysis (MLDA) for face localization and lip detection and lip reading problems [19]. Puviarasan et al. presented a system starts with detecting the face region and then the mouth is detected relatively. Next, the mouth features are extracted using DCT and discrete wavelet transform (DWT) and they are recognized using the HMM [20].

3. ”Real time” multi-elements AVSR systems

Recently, specifically since 2011, the offline multi-elements AVSR systems have received much interest towards development in the real time. In principal, real time means that the elements (face/mouth detection, face recognition, word recognition, etc.) are all performed online. According to the best of our knowledge, there are only two real time AVSR systems have been developed in 2011. The first system is developed by Shin et al. who presented a multimodal AVSR system for Korean language [21]. Shins system includes face detection, eye detection, mouth detection, mouth tracking via mouth end-point detection, and mouth fitting. All these are achieved using a package of multiple methods, namely, AAM, LucasKanade feature tracker, fast block matching algorithm and artificial neural network (ANN). Then they used three different classifiers HMM, artifical neural network (ANN), and k-nearest neighbor (k-NN). The other real time system is developed by Saitoh et al. who presented a real time VSR system for Japanese language [22]. Although the basis of Saitohs system is a method already proposed by him before, his new system adopted the user facial movements to enhance the VSR system.

AVSR and Humanoid Robots

Most of the traditional systems, described in section 2.1, are designed to be used indepedently. However in recent years, specifically since 2008, and with the great development of HRs, there is an urgent need to combine AVSR systems with HRs. In literature, there are number of articles, treat HRs; use the AVSR as the communication pipeline. In [23] Hao et al. claimed that the integration of AVSR system provides more natural and friendly human computer interaction (HCI). They presented an approach leads to a text-driven emotive AV avatar. In [24] Guan et al. discussed the challenges in the design of HRs through two types of examples; emotion recognition and face detection. Their AV system is based on bimodal emotion recognition and face detection in crowded scene. In [25] Yoshida et al. proposed a two-layered AV integration framework that consists of AV voice activity detection based on a Bayesian network and AVSR using a missing feature theory to improve performance of ASR. The ASR system is implemented with the proposed two-layered AV integration framework on open-source robot audition software denoted as HARK. In [26] Chin et al. presented an AVSR system combines three modules RBF-NN voice activity detection, watershed lips detection and tracking and multi-stream AV back-end processing.

3 Motivation and work outline

As the above short review explained, most of traditional AVSR systems are concerned only with the lip movements and acoustic stream. They neglect the element of UID which, certainly, enhances the recognition of the user personality. Additionally, the traditional systems force the user to fix himself/herself in front of the camera and avoid any possible movements, which cannot be avoided in modern applications. Furthermore, the computations of most of the above systems are complex since they include the usage of several techniques which consumes much time and contradicts with the concept of real time. These aspects represent real obstacles in the way of developing robotics applications based on audio-visual speech. As it dipcted in Figure 2, the system outline includes two main parts, visual part and audio part. In the visual part, the user appears in front of the camera, the user face is detected and, then, the mouth is localized. Then, the user name is identified and written on the program console. Once the user utters a word, the visual frames, which include lip movements, are saved in a folder and insert to be the input of SOM and then k-NN for classification. In the audio part, the acoustic stream of the uttered word is saved in a folder and insert to be the input for MFCC and then HMM for classification. A late integration step is performed between visual cue and audio cue. Using a weighted probabilistic rule, the word is recognized and written on the console. The

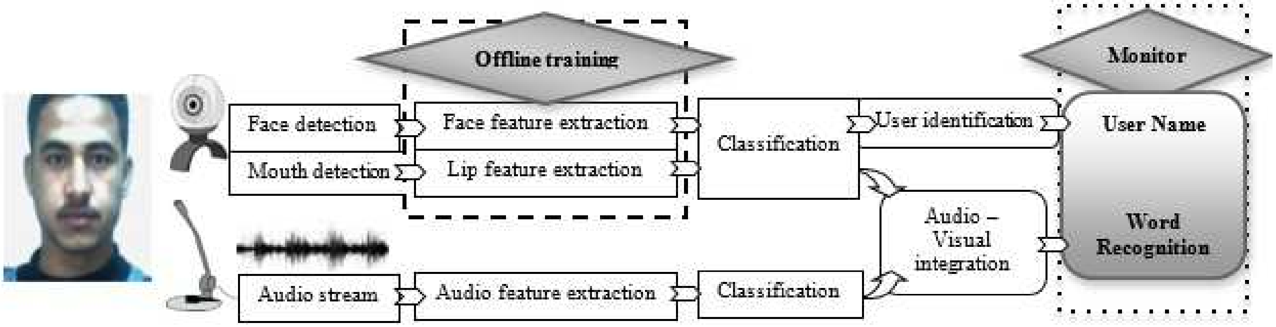


Fig. 2: The proposed system including the three modules. Dashed square encloses offline training and the dotted one encloses monitor.

contributions of the proposed system are many. It is clear that, there are no restrictions on the movements of the user as he/she in the scope of the camera. Also, the user identification element distinguishs the proposed system from other traditional systems. Additionally, the system computation is simple, where it uses only four techniques in one code environment. Furthermore, the lip localization algorithm presented in this paper is new and recovered the problem of localization of Viola-Jones algorithm [27], which we adopted in our work [2]. Finally, the current paper describes the first audio-visual database uses Arabic language.

4 VSR system overview

In [2], the author presented a VSR system includes three components:

Face and mouth detection

The VSR system [2] adopted the Viola-Jones detection module [27], which is much faster than any of its contemporaries via the use of an attentional cascade using low feature number of detectors based on a natural extension of Haar wavelets. In this cascade, each detector fits objects to simple rectangular masks. In order to reduce the number of computations for such large number of cascades, Viola and Jones used the concept of integral image. They assumed that, for each pixel in the original image, there is exactly one pixel in the integral image, whose value is the sum of the original image values above and to the left. The integral image can be computed quickly, which drastically improves the computation costs of the rectangular feature models. The attentional cascade classifiers are trained on a training set as the authors have explained in [27]. As the computation progresses down the cascade, the features can get smaller and smaller, but fewer locations are tested for faces until detection is performed. The same procedure is adopted for mouth detection, except that object is different and search about mouth will be only on the lower half of the input image. Figure 3 shows the real time face and mouth detection. *(Please pay attention that, an enhancement to lip detection is presented in this paper, see section 5.4)*

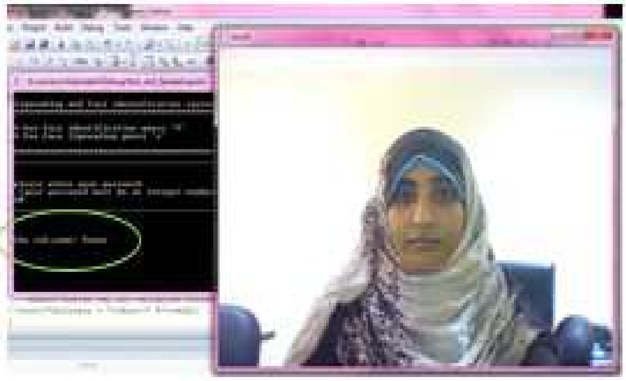


Fig. 3: (Frame) there is a square around facial features and a rectangle around mouth. (Console) the oval shape encloses the UID

User identification (UID)

User identification element distinguishes our system than other systems; see recent examples [18,19,20,21,22]. We believe that the UID element is so important, especially in HRI applications, where the user identity is important for security restrictions and authentic communications. The system shows the ID of the user once his/her face is detected, see Figure 3. This is done automatically, where the system saves one frame for the detected face and then two tasks are performed on this frame: feature extraction and, then, feature recognition. The first task is achieved using SOM [28]. The SOM is a well-known artificial neural network applies unsupervised competitive learning approach. In each learning step, one sample input vector I from the input data is considered and a similarity measure, usually taken as the Euclidian distance, is calculated between the input and all the weight vectors of the neurons of the map. The best matching neuron (BMN) c, called winner, is the one whose weight vector wc has the greatest similarity (or least distance) with the input sample I; i.e. which satisfies:

$$\|I - w_c\| = \min_u(\|I - w_u\|), \tag{1}$$

After deciding the winner neuron, the weights of the map neurons are updated according to the rule:

$$w_u(t+1) = w_u(t) + h_{cu}(t)[I(t) - w_u(t)] \tag{2}$$

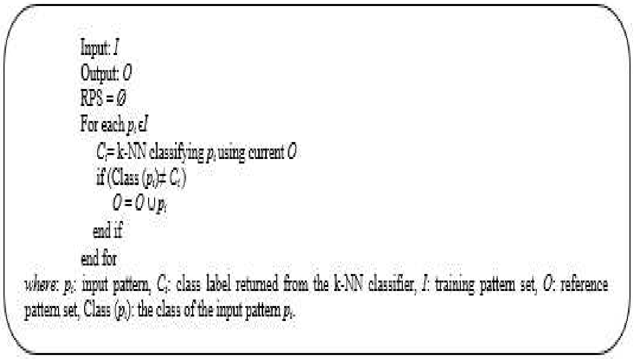
where

$$h_{cu}(t) = \alpha(t) \times exp(\frac{r_c - r_u}{2\sigma^2(t)}) \tag{3}$$

$h_{cu}(t)$ is the neighborhood kernel around the winner c at time t, $\alpha(t)$ is the learning rate and and is decreased gradually toward zero and $\sigma^2(t)$ is a factor used to control the neighborhood kernel. The term $r_c - r_u$ represents the difference between the locations of both the winner neuron c and the neuron u. After training phase, the neuron sheet is automatically organized into a meaningful two-dimensional order map denoted as a feature map (or codebook). The SOM codebook has the merit of preserving the topographical order of the training data. In other words, similar features in the input space are mapped into nearby positions in the feature map [28]. *(Please note that, a new search algorithm for SOM is used in this paper, see section 5.2).* Regarding to the feature recognition task, we used k-NN classifier [29]. The k-NN is a well-known non-parametric classifier has a simple structure and exhibits effective classification performance, especially, when variance is clearly large in the data, as our situation here. The k-NN classifier has a labeled reference pattern set (RPS) for each class being determined during training phase. The input of k-NN, i.e. the observed features, is the weight vector of the winner neuron of SOM, which will be compared with each reference feature (the SOMs face codebook) using the Euclidian distance. Then, we choose the set k of nearest neighbors and determine the class of the input feature using a majority voting procedure. To construct the best k-NN classifier, it is not practical to use all training sample data as the reference pattern set. Instead, we constructed the k-NN using Harts condensing algorithm [30], which effectively reduces the number of RPSs. Figure 4 shows the k-NN algorithm used in this paper.

Visual speech recognition VSR

The VSR system was the main contribution presented in [2], where once the system identifies the user, then the user starts to utter a word. The system captures the visemes included in this word across number of lip frames. Then, by the same way described in UID module, two tasks are performed on these frames: viseme features extraction and, then, viseme features recognition. In the first task, we reused the SOM again by the same way given in equations (1)-(3), except that the object here is the mouth region. In the second task, also we used the k-NN classifier by the same way described in Figure 4.



```
Input: $I$
Output: $O$
RPS = $\emptyset$
For each $p_i \epsilon I$
    $C_i$= k-NN classifying $p_i$ using current $O$
    if (Class ($p_i$)$\neq C_i$)
        $O = O \cup p_i$
    end if
end for
```

where: p_i: input pattern, C_i: class label returned from the k-NN classifier, I: training pattern set, O: reference pattern set, Class (p_i): the class of the input pattern p_i.

Fig. 4: The k-NN training procedure using Harts condensing algorithm [30]

5 Enhancement of the VSR system

In this paper, we adopt some modifications on the VSR system described in [2] according to our observations during experiments. These observations led, in our opinion, to the degradation of the VSR systems performance. This section addresses these observations and our approach to recover for the sake of improving the overall AVSR performance.

Preprocessing

During the experiments of VSR system described in [2], we noticed that there is degradation in its performance. In our opinion, this degradation dues to the bad lightning conditions. Bad lightening conditions make the performance of Viola-Jones detection module something inefficient. As such, k-NN could not classify all words correctly, which causes the degradation of the overall recognition rate. To recover this problem, some kind of preprocessing for the input images is needed in advance. This preprocessing step includes the normalization of the input images by making contrast stretching for each image. Contrast stretching is a kind of normalization enhances the images by attempting to improve the images contrast by stretching the range of intensity values it contains to span a desired range of values. Specifying the upper and the lower pixel value limits over which the image is to be normalized can perform this stretching. Often these limits will be the minimum and maximum pixel values that the image type are concerned. For example, for 8-bit gray level images the lower and upper limits might be 0 and 255. We can call the lower and the upper limits a and b respectively. The simplest way of normalization then scans the image to find the lowest and highest pixel values currently present in the image, for example c and d. Then each pixel P is scaled using the following rule:

$$P_{out} = (P_{in} - c)(b - a/d - c) + a \tag{4}$$

Values below 0 are set to 0 and values about 255 are set to 255. When we applied this rule in new experiments, we found another problem with this technique. The problem is that a single outlying pixel with either a very low or very high value can severely affect the value of c or d and this could lead to very unrepresentative scaling. Therefore, we used a more robust approach for deciding c and d. We take the histogram of each image, and then select c and d at, for example, the 5th and 95th percentile in the histogram (that is, 5 % of the pixel in the histogram will have values lower than c, and 5% of the pixels will have values higher than d). Experimentally, we found that this approach prevents the outliers from affecting the scaling so much. In the following, Figure 5 shows the representation of more than 4000 lip images before and after achieving the normalization approach given in this section. There are two panels, the upper is before normalization and the lower is after normalization. It is clear that before normalization, there is a clear scattering for the data as a result for the existence of bad illumination conditions. The oval shape encloses a wide scattered area, whereas the arrow explains the width of the data distribution. After normalization, data representation is more condensing than before normalization. The oval shape became smaller and includes less number of points and the arrow became shorter.

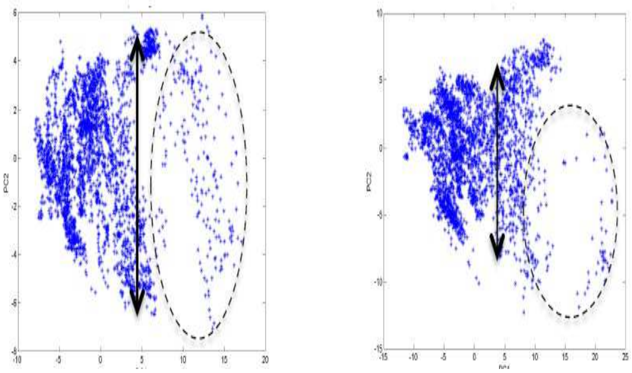


Fig. 5: Data (up) before normalization (bottom) after normalization. Oval shape & arrows refer to data-condensing

Mouth detection

Another problem stems from using the detection module of Viola-Jones [27] in detecting mouth is that the box drawn around the mouth disappears when the user opens, unintentionally, his/her mouth a little bit larger. In such case, the sequence of the lip movements is broken which causes the loose of information included in these frames. We check this phenomenon by following the performance of k-NN in such situations; we found that k-NN usually misclassified the words that include such phenomenon. Accordingly, in this paper we replace the mouth detection module described in section 4.1 with a mouth localization algorithm based on deciding the bounding box around the mouth in a geometric way. Proportional to the bounding box around the face, the algorithm decides first the left corner point of the mouths bounding box. Then, the required size of this box can be drawn easily to the extent encloses any possible lip movements. Table **??** shows the proposed mouth localization algorithm. Figure **??** (a) and (b) show how the algorithm decides the box around the mouth in proportion to the box around the face. The impact of the mouth localization algorithm is clear in Figure **??** (c) and (d). For the same image, in (c) the box is drawn by the same way given in [2] whereas in (d) the box is drawn using the algorithm in Table **??**. It is clear that the box around the mouth in the second case is larger than the box of the first case.

P1D-SOM as a feature extractor

In this paper and for the sake of real time experiments, instead of using the traditional algorithm of SOM we use a fast search algorithm denoted as P1D-SOM [31] presented by the author. The idea of P1DSOM is based on the idea of principal components of the feature space during the recognition or testing phase. Instead of calculating the distance between the input image and all neurons in the trained feature map, calculate the distance between the input image and only the neurons lie on the principal components of the feature map, especially when we used an SOM with more than two dimensions. By this way, P1DSOM reduces, drastically, the computation time of SOM during the testing phase which enables the real time processing. The abilities of P1D-SOM are appeared clearly when large databases are used, which require high

Table 1: Mouth localization algorithm

```
1.Grab the video frame for input,
2.Achieve the face detection and draw a box on the detected face then determined some of detection box (face) properties where:
  •The origin point
    –$X_f$: x-coordinate of the left border of face region
    –$Y_f$: y-coordinate of the top border of face region
  •The Width
    –$W_f$: the width of face region
  •The height
    –$H_f$: the height of face region
3.Detect the lip region is set as per the following calculations,
    –$X_l = X_f + W_f/4$,
    –$Y_l = Yf + (2*Hf/3)$,
    –$Wf = Wf/2$,
    –$Hl = Hf/3$
  where
    –$X_l$ : x-coordinate of the left border of lip region
    –$Y_l$ : y-coordinate of the top border of lip region
    –$W_l$ : the width of lip region
    –$H_l$ : the height of lip region
4.$X_l, Y_l, W_l$ and $H_l$ are the values constituting of the lip region in lip detection.
5.Repeat steps 2, 3 and 4 for all frames.
```

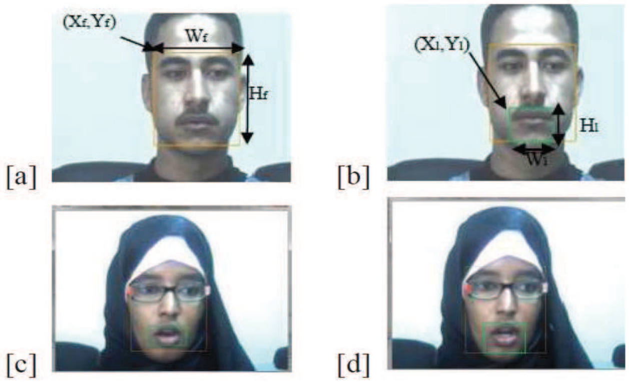


Fig. 6: The mouth localizing algorithm (a) the box around face (b) the box around mouth in proportional to the face box (c) the box around mouth using Viola-Jones module given in [2] (d) the box around mouth using the algorithm described in Table **??**.

dimensional feature map, more than two dimensions. In this paper, for simplicity, we will refer to P1D-SOM as SOM. For more details about P1DSOM please refer to [31].

HMM Classifier

However the k-NN performance as a classifier for the VSR system described in section 4 is good enough, in this paper we tried to insert the SOM features to the HMM in order to do the same job of k-NN and, finally, compare between both performances. It is known that, HMM holds the greatest promise among various tools used for automatic speech recognition studied so far due to its capabilities in handling either the sequence or the variability of speech features [32]. For isolated word recognition, each word is modeled by HMM model. HMM word models are constructed with the sub-states of Gaussian mixture model and the state transition probability. Therefore, we must pre-define the number of states for each word, and the number of Mixture of Gaussian (MoG) for each state. In the training process of HMM, it is easy to obtain the number of MoGs of each state and the transition probability of each state as well, however, in the classification process we search for the best state transition path and determine the likelihood value using the Viterbi search algorithm [32]. We compare the likelihood value of each HMM and determine the input word as the index of HMM with the largest likelihood value.

6 Audio-Visual Speech Recognition (AVSR)

As Figure 3 shows, once the user is identified and his/her mouth region is detected this launches the second part in the proposed AVSR system. The AVSR includes two components, the ASR and the VSR. The performance of the AVSR is highly dependent to three main factors: (1) performance of the ASR, (2) performance of the VSR, and (3) effectiveness of the AV integration. In the experiments of AVSR, both the audio and visual data streams are first obtained from the input speech and pumped into the respective module for further processing. We already described the visual component VSR, in the following, we describe the ASR component and AV integration.

Audio speech recognition (ASR)

The input audio stream takes the direction of audio speech channel, which is shown in the bottom part in Figure 2. Through the audio channel, the ASR module performs the tasks of feature extraction and, then, feature recognition on the audio stream.

Audio speech recognition (ASR)

Once the audio stream is gathered, the audio features are extracted using one of the most representative methods for audio speech, which is the Mel Frequency Cepstral Coefficient (MFCC) [26]. In the MFCC, the audio input signal is first windowed into a fixed length size. Then, the windowed signal is pre-emphasized for spectrum flattening purpose using the Finite Impulse Response (FIR) filter [33], as follows:

$$H(z) = 1 - az^{-1} \tag{5}$$

where a is normally chosen to be a value between 0.9 to 0.95. Once framing is done, each frame goes to for the Hamming filtering in order to smooth out the signal. In speech analysis domain, Hamming function is one of the most commonly used window functions capable to reduce the spectral distortion and signal discontinuities in the speech frames [34]. After we apply the Hamming window as the aforementioned, the output signal will be represented as follows:

$$h(n) = g(n)w(n) \tag{6}$$

where $0 \leq n \leq N-1$, g(n) = segmented voiced signal and h(n) = output windowed signal. Next to performing the process of framing and windowing for each frame, fast Fourier (FFT) is used in order to convert the frame data from time domain into frequency domain. The FFT for the windowed signal, which consists of length N, could be represented mathematically as follows [33]:

$$X(k) = \sum_{j=1}^{N} x(j) w_N^{(j-1)(k-1)} \tag{7}$$

where $w_N = \mathrm{e}^{-2\pi i}/N$ is the N^{th} root of the unity. The decibel amplitude of the spectrum, which is the output from FFT, is then mapped onto the Mel scale, which is formed according to the concept the frequency of 1 kHz is equal to the Mel-frequency of 1000 Mels. It is known that, the relationship between Mel-frequency and the frequency is given as follows [26]:

$$M(j) = \sum_{n=1}^{N} log(|mt(mel)|) cos(k\pi/N)(n-0.5) \quad k = 0, \dots, j \tag{8}$$

After performing the frequency wrapping using the Mel-frequency filter bank, the log Mel-scale is converted back into the time domain by using the discrete cosine transform (DCT). Then, the extracted audio features are saved in the form of a vector and passed to the recognizer.

Audio feature recognition

For this task, we reuse the HMM as a recognizer by the same way described in section 5.4 but the signal here is audio not video.

Audio visual speech integration (AVSI)

In literature, there are two main different architectures are provided for AVSI: feature-based integration and decision-based integration [26]. Feature-based integration combines the audio features and visual features into a single feature vector before going through the classification process, for such case, one classifier is enough. In the decision-based integration, two parallel classifiers would be involved rather than one. In this paper, we adopted the latter architecture, where we used the traditional HMM for audio features classification and k-NN for visual features classification. The result from each classifier is fed into the final decision fusion, such as probabilistic weighted basis, to produce the final decision of the corresponding word.

7 Database

In this paper, we use the audio-visual Arabic speech (AVAS) database developed by the author [35] and considered as an extension to our database described in [2]. AVAS extended the number of words to be 36 words instead of 9 words only in [2]. Also, AVAS includes 13 Arabic phrases where previous database [2] did not include phrases. Additionally, the number of subjects in AVAS is increased to be 50 subjects instead of 20 subjects in [2]. Of course increasing the number of subjects to be 50 persons enlarges the included information in the database and ensures the generalization of our system toward variation of users. According to the best of our knowledge, AVAS database is the first audio-visual Arabic speech database presented so far in literature [35]; see samples of subjects in Figure 7. Table 2 includes the 26 Arabic words used in the experiments of this paper, where each word includes between 2-4 phonemes. The word list includes Arabic digits (1,2,3,9), (10,20,30,100) and days (Saturday, Sunday,, Friday) all in Arabic language.

However a separate article includes AVAS is already submitted and the database itself will be free for public use shortly, we show here some details about it. We captured the AVAS video data with 640 x 480 pixel resolution at 30 fps frame rate and saved in AVI format. Each word/person spans a separate video. But for the sake of training and testing the word classifiers, we break each video into 20 frames per word. Each frame is cropped to have a 48x48 pixel resolution. Since the database has the recording of 50 subjects, then in total, AVAS includes (50 persons x 26 words x 20 frame) 26,000 images. The AVAS audio data is recorded using a USB microphone attached to the PC, where each video includes, synchronously, the acoustic stream of the uttered word. This audio stream is saved later in a separate wav file. The AVAS database is gathered in the image processing laboratory at Center for Artificial Intelligence and RObotics (CAIRO) in a general office environment with different illumination conditions. Figure 8 shows an experimental session.

The 50 subjects included in the database are mix of female and male. Most of the females wear the Arabian scarf on their heads, whereas few of the male wear the traditional hair cover. Usually, the female scarf may hide some parts from the face such as hair, ears, forehead, parts of eyes, and parts of cheek. However most of these parts represent important information, we keened to make the database natural. So we adapt with this point by adjusting the feature extractor. The real-time experiments are implemented on a PC with an Intel(R) Core2Duo CPU at 2.39 GHz, 2 GB RAM and a Microsoft Win7 Professional (32bit) operating system. We used the run-time program using Microsoft Visual C++ 8.0. For the video capture, we used a Pixelink USB Cam version (PL B762U) and (Golden Video software program) to capture and control the video.

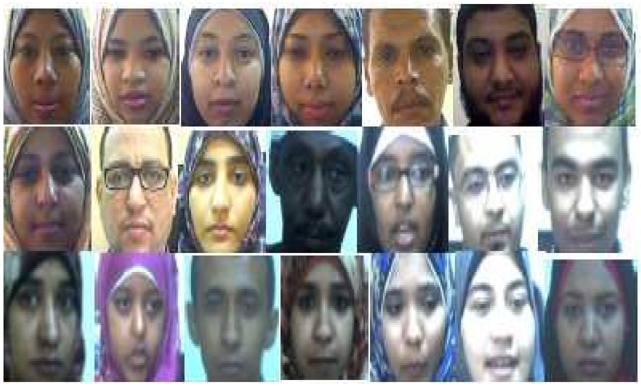


Fig. 7: Samples of the database.

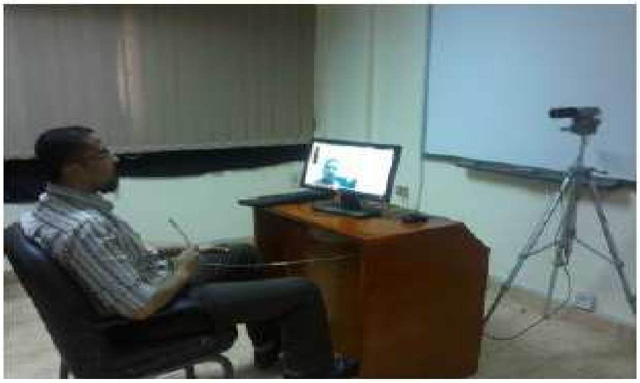


Fig. 8: An experimental session.

Table 2: Word correct rate (%) of the proposed system

Word code	Arabic word	Pronunciation	English meaning
1	وَاحِد	/wa-he-d/	One
2	اثنين	/et-nee-n/	Two
3	ثلَاثة	/ta-la-ta/	Three
4	اربعة	/ar-ba-aa/	Four
5	خمسة	/kha-m-sa/	Five
6	سته	/se-taa/	Six
7	سبعة	/sa-ba-aa/	Seven
8	ثمَانية	/ta-ma-nya/	Eight
9	تسعة	/te-se-aa/	Nine
10	عشره	/a-sh-raa/	Ten
11	عشرين	/e-sh-ree-n/	Twenty
12	ثلَاثين	/ta-la-tee-n/	Thirty
13	اربعين	/ar-ba-ee-n/	Forty
14	خمسين	/kha-m-se-n/	Fifty
15	ستين	/se-tee-n/	Sixty
16	سبعين	/sa-b-ee-n/	Seventy
17	ثمَانين	/ta-ma-nee-n/	Eighty
18	تسعين	/te-se-ee-n/	Ninety
19	مَائة	/me-aa/	Hundred
20	سبت	/sa-ba-t/	Saturday
21	احد	/a-ha-d/	Sunday
22	اثنين	/et-nee-n/	Monday
23	ثلَاثَاء	/thu-la-th-aa/	Tuesday
24	اربعَاء	/ar-ba-aa/	Wednesday
25	خميس	/kha-m-ee-s/	Thursday
26	جُمعة	/gu-ma-aa/	Friday

8 Experimental Results

Experimental setup

For the experiments of this paper, we built two separate feature SOM map (or codebook), one for the UID module and the other for the VSR module. For UID map, it includes 18x14x4 neurons in three dimensions. To build this map, i.e. during training phase, we just used ten samples for each subject to be in total 500 input images. For the testing phase, every subject tries the system 5 times in different sessions (sometimes different days). Then the total number of trials is: 50 (subjects) x 5 (trials) = 250 trials. For the VSR experiments, the SOM map includes 28x24x20 neurons in three dimensions. We used the samples of 25 subjects for training, i.e. 25 (subjects) x 26 (words) x 20 (frames) = 13000 input images, and the samples of the other 25 subjects, same number of training images, are reserved for testing. The AVSR experiments are running through two phases, the first is person dependent phase (PD) where the subjects used in training sessions are used also in testing sessions. The second phase is person independent (PI), where the subjects used in testing experiments are totally new and did not used in training experiments. Certainly, this ensures the generalization of the proposed system. When the user starts to utter a word, the system saves the frames of the uttered word in a folder and the audio stream in a wav file. The frames are applied to be the input of SOM, which starts to extract the words visemes or features of the word. Then, SOM sends the extracted feature in the form of a vector to the classifier; k-NN or HMM. The classifier compares the coming vector with the SOMs visemes codebook, which we got after training and yield a result. The wav file, which includes the audio stream, is applied as the input to MFCC for audio feature extraction and then audio feature recognition using HMM and yield another result. The result from each classifier is fed into the final decision fusion, a probabilistic basis, to yield the final decision of the corresponding word. Regarding classifiers, please pay attention that we determined the optimal structure of each word classifier experimentally. In the case of HMM-based word classifier, the extracted features are used to create an HMM model for each word. Each HMM model is a 5-state left to right model with 3 Gaussian mixtures for each (there is a possibility to adapt this number in the future). The HMM models are initialized and subsequently re-estimated with the embedded training version of the BaumWelch algorithm. Then, the training data was aligned to the models via using the Viterbi algorithm to obtain the densities of the state duration. The maximum probability model is recognized as the output word model and the corresponding word is displayed on the systems console in the form of a text. In the case of k-NN based word classifier, the number of nearest neighbors k is chosen to be 3.

Description of the results

User identification results

For the UID experiments, the system achieved user identification successfully in 241 trials out of the total 250 trials, i.e. identification rate is about 96%. Please note that, real time (or testing) experiments were run in different days and sessions than those of training sessions. During the testing experiments, we noticed that some of the misidentification cases occur in cases of female who wear scarf. Wearing scarf with possible bad lightening conditions may confuse the classifier. However, we believe that repeating and adjusting the training phase of SOM ensures to recover this problem.

Visual speech results

For the VSR experiments, in the PD phase, we asked the 25 subjects, who used to train the system, to test the system in real time experiments. In other words, the users to the system are not new. We found that the combination SOM+k-NN outperforms the combination SOM+HMM. Namely, the first combination achieves 85.7% where the other combination achieves 80.4%. In the PID experiments, we asked the other 25 subjects, who are different than those used in training phase, to test the system in real time sessions. Again, the SOM+k-NN gives 61.1 % whereas the SOM+HMM gives 57.4 % as recognition rate. These results are included in Table 3.

Table 3: Word correct rate (%). VSR-Comparison between HMM and k-NN classifiers

	Visual Speech Recognition (VSR) (%)	
Type	SOM+k-NN	SOM+HMM
PD phase	85.7	80.4
PID phase	61.1	57.4

Audio speech and AVSR results

For the ASR experiments, in the PD phase, we got 87.3% whereas in the PID experiments we got 65.4% as recognition rate. For The AVSR experiments, we got 94.1% whereas in the PD experiments we got 70.3% as recognition rate, see table 4. Please pay attention that, the big difference between results of PD phase and PID phase is natural, since in the PD phase, the subjects of training and testing are the same. However, in the PID phase the subjects used in training are different than those used in testing. Therefore, PD phase does not include a challenge for the classifier. But, needless to remind that PID experiments are more important and general than PD experiments since the PID experiments measure and indicate to the system generalization. The most important to say here is that the combination between audio signals with visual cues enhances the performance of each. In PD phase, the VSR alone achieves 85.7%, the ASR alone got 87.3% and the AVSR achieves 94.1%. In the PID phase, the VSR alone achieves 61.1%, the ASR alone got 65.4% and the AVSR achieves 70.3%. Therefore, the AVSR performance is better than both the audio alone and visual alone speech recognition.

Table 4: Word correct rate (%)-ASR and AVSR of the proposed system using k-NN

Word Correct Rate (WCR) (%)		
Type	ASR	AVSR
PD phase	87.3	94.1
PID phase	65.4	70.3

Comparison with another reported system

According to the best of our knowledge, this is the first AVSR system use Arabic language database is designed so far. So we couldnt compare the proposed system performance with any similar system. On the same time, conducting a comparison with other reported systems utilizing different databases is not a fair way for judgment as well. However, we can do this comparison only as evaluation for our systems performance. Table 5 shows a comparison with another reported real time system for isolated Korean words [21] in PID phase only, where it is easy to remark that the proposed system outperforms the system in [21] in both VSR and AVSR modules.

Table 5: Comparison with another real time system [21]

Approach	Proposed	System [21]
VSR	61.1	46.5
AVSR	70.3	60.0

Beside the recognition rate, we can determine more advantages of the proposed system over the system [21]. First of all, in [21] the number of subjects is fourteen, thirteen subjects were used for training and only one used for testing. In fact, we could not assume that a system can generalize well using only one subject. Second, the authors in [21] used many techniques in order to achieve their system. Namely, they used adaptive boosting (AdaBoost) algorithm to achieve face detection, eye detection and mouth detection, then they applied two detection methods: mouth-end points to the Active Appearance Method (AAM) and then apply AAM fitting algorithm to the mouth area. For lip tracking they used model-based LucasKanada algorithm for tracking outer lip and fast block matching algorithm for tracking inner lip. Next, they used neural network classifier to achieve lip detection activation. Finally, they used three kinds of classifiers separately for word recognition. We believed that the system [21] passes many phases or utilized many techniques in order to achieve its AVSR module. Indeed, these multiple phases added complexity to the final system. In contrast, our system here is simple where we used a simple detection module. Then we used SOM and k-NN two times, using a recursive call, one for UID and the other for VSR. For ASR we combine MFCC with HMM. So, in general, the computation complexity is much less than the Korean system [21].

9 Conclusion & future work

Based on the visual speech recognition (VSR), or lip reading, system presented by the author in [2], we proposed here a valuable upgrade and enhancement for this VSR system. Here we presented few steps to enhance the system described in [2]. Also, we combined the audio signals to the visual signals in order to produce an overall AVSR system works on the real time. Unlike traditional offline systems, the proposed system adapts to the user facial movements that cannot be avoided in a real life. For the sake of real time, we used a fast search algorithm for the Self Organizing Map (SOM) in order to achieve feature extraction task. SOM reduces the high dimensional input space into low dimensional feature space. For the classification task, we used k-Nearest Neighbor (k-NN) and Hidden Markov Model (HMM) in order to recognize the features extracted by SOM. In the experiments of this paper, we increased the number of subjects to be 50 persons instead of 20 and the number of words to be 26 instead of 9 in [2]. Increasing the number of subjects and words confirms the generalization of our system. Experimental results showed that the proposed AVSR system is promising. With more adjusting for the SOM training phase, the testing results will be improved drastically. Also, the system is comparable where the comparison with another reported system showed that the proposed system is better and simple. The proposed AVSR system can be combined easily with any humanoid robot as a multimodal interaction module or used alone. For future works, we will adjust both of the number of each HMM models state and the value of k in k-NN. In addition, we will upgrade the presented system to achieve audio visual Dialogue as a multimodal interaction with the computer to produce a health care agent based Arabic language.

Acknowledgement

The presented work is funded by Science and Technology Development Fund (STDF) program, Ministry of State for Scientific Research, Egypt via the research grant no. 1055.

References

[1] G. Bailly, P. Perrier, V-B. Eric, . Audiovisual Speech Processing, Cambridge University Press.

[2] A. Sagheer, S. Aly, Integration of Face Detection and User Identification with Visual Speech Recognition. In Proceedings of the 19th International Conference On Neural Information Processing. ICONIP '12, In T. Huang et al. (Eds.): ICONIP 2012, Part V, LNCS 7667, 479-487. Springer (2012).

[3] G. Potamianos, Audio-Visual Speech Processing: Progress and Challenges. In Proceeding of the HCSNet Workshop on the Use of Vision in HCI. VisHCI '06, Conferences in Research and Practice in Information Technology (CRPIT), 56, R. Goecke, A. Robles-Kelly & T. Caelli, Eds (2006).

[4] G. Potamianos,C. Neti, G. Gravier, A. Garg, A. Senior, Recent advances in the automatic recognition of audiovisual speech. In Proceedings of the IEEE, **91**, 9, 1306-1326 (2003).

[5] J. Luettin, N. Thacker, Speech reading using Probabilistic Models. Computer Vision and Image Understanding, **65**, 2, 163-178 (1997).

[6] S. Dupont, J. Luettin, . Audio-Visual Speech Modeling for Continuous Speech Recognition. IEEE Transaction on Multimedia, **2**, 3 (2000).

[7] I. Matthews, G. Potamianos, C. Neti, J. Luettin, A comparison of model and transform-based visual features for audio-visual LVCSR. In Proceeding of IEEE International Conference on Multimedia and Expo. ICME '01, 825 828 (2001).

[8] I. Matthews, T. Cootes, A. Bangham, S. Cox, R. Harvey, Extraction of Visual Features for Lip-Reading. IEEE Trans. on Pattern Analysis and Machine Intelligence, **24**, 2 (2002).

[9] M. Heckmann, K. Kroschel, C. Savariaux, F. Berthommier, DCT-Based Video Features For Audio-Visual Speech Recognition. In Proceedings of Inter. Conf. on Spoken Language Processing. ICSLP '02, 1925-1928 (2002).

[10] P. Yin, I. Essa, J.M. Rehg, Asymmetrically Boosted HMM for Speech Reading. In Proceedings of the IEEE Conf. on Computer Vision and Pattern Recognition. CVPR '04, **2**, 755-761 (2004).

[11] J.F. Guitarte, A.F. Frange, E.L. Solano, K. Lukas, Lip Reading for Robust Speech Recognition on Embedded Devices. In Proceedings of the 30th IEEE Int. Conf. on Acoustics, Speech, and Signal Processing. ICASSP '05, vol1, 473 476 (2005).

[12] T. J. Hazen, Visual Model Structures and Synchrony Constraints for Audio-Visual Speech Recognition. IEEE Transaction on Speech and Audio Processing, **14**, 3, 1082-1089 (2006).

[13] A. Sagheer, N. Tsuruta, R. Taniguchi, S. and Maeda, Appearance Features Extraction vs. Image Transform for Visual Speech Recognition. International Journal of Computational Intelligence and Applications, IJCIA, 6 (1), 101-122, Imperial College Press ICP (2006).

[14] A. Sagheer, N. Tsuruta, R. Taniguchi, S. Maeda, S. Hashimoto, "A Combination of Hyper Column Model with Hidden Markov Model for Japanese Lip-Reading System," Proc. of the 4th International Symposium on Human and Artificial Intelligence Systems, HART04, Japan (2004).

[15] A. Sagheer, N. Tsuruta, R. Taniguchi, S. Maeda, Hyper Column Model vs. Fast DCT for Feature Extraction in Visual Arabic Speech Recognition, Proc. of the 5th International IEEE Symposium on Signal Processing and Information Technology, ISSPIT05, 761-766, Athens, Greece (2005).

[16] C. Sanderson, K. Paliwal, Identity verification using speech and face information. Digital Signal Processing, **14**, 449480 (2004).

[17] H.E. Cetingl, E. Erzin, Y. Yemez, A.M. Tekalp, Multimodal speaker/speech recognition using lip motion, lip texture and audio. Signal Processing, **86**, 12, 3549-3558 (2006).

[18] T. Saitoh, K. Morishita, R. Konishi, Analysis of efficient lip reading method for various languages. In Proceedings of the 19th International Conference on Pattern Recognition ICPR, 14 (2008).

[19] M. Deypir, S. Alizadeh, T. Zoughi, R. Boostani, Boosting a multi-linear classifier with application to visual lip reading. Expert Systems with Applications, **38**, 1, 941-948 (2011).

[20] N. Puviarasan, S. Palanivel, Lip reading of hearing impaired persons using HMM. Expert Systems with Applications, **38**, 4477-4481 (2011).

[21] J. Shin, J. Lee, D. Kim, Real-Time Lip Reading System for Isolated Korean Word Recognition. Pattern Recognition, **44**, 559-571(2011).

[22] T. Saitoh, R. Konishi, Real-Time Word Lip Reading System Based on Trajectory Feature. The IEEJ Transactions on Electrical and Electronic Engineering, **6**, 289-291 (2011).

[23] T. Hao, F. Yun, T. Jilin, H-J and Mark, Humanoid audio-visual avatar with emotive text-to-speech synthesis. IEEE Transactions on Multimedia, **10**, 969-981 (2008).

[24] L. Guan, W. Yongjin, T. Yun, Toward natural and efficient human computer interaction. In Proceedings of the IEEE International Conference on Multimedia and Expo. ICME '09, 1560-1561 (2009).

[25] T. Yoshida, N. Kazuhiro, G. O. Hiroshi, Automatic speech recognition improved by two-layered audio-visual integration for robot audition. In Proceedings of the 9th IEEE-RAS International Conference on Humanoid Robots (Humanoids 2009), 604-609 (2009).

[26] S.W. Chin, K.P. Seng, L. Ang, Audio-Visual Speech Processing for Human Computer Interaction. In T. Gulrez, A.E. Hassanien (Eds.): Advances in Robotics & Virtual Reality, ISRL 26, 135-165, Springer (2012).

[27] P. Viola, M. Jones, Robust real-time object detection. The IEEE Transactions on Computer Vision, 57 (2), 137-154 (2004).

[28] T. Kohonen, Self-Organizing Maps, the 3rd Edition, Springer (20101).

[29] W. Xindong, others, Top 10 algorithms in data mining. Knowledge Information Systems, **14**, 1-37 (2008).

[30] P. Hart, The condensed nearest neighbor rule. The IEEE Transaction on Information Theory, 14, 515-516 (1968).

[31] A. Sagheer, N. Tsuruta, R. Taniguchi, P1DSOM- A Fast Search Algorithm for High-Dimensional Feature Space Problems. International Journal on Pattern Recognition and Artificial Intelligence, IJPRAI, World Scientific, **28**, 2, 1459005 (2014).

[32] L. Rabiner, A tutorial on hidden Markov models and selected applications in speech recognition. In IEEE Proc., **77**, 136-257 (1989).

[33] P. Denbigh, System analysis and signal processing with emphasis on the use of MATLAB. Addison Wesley Longman Ltd (1998).

[34] Y. Song, X. Peng, Spectra Analysis of Sampling and Reconstructing Continuous Signal Using Hamming Window Function. In Proceedings of the 4th International Conference on Natural Computation (ICNC 2008) (2008).

[35] S. Antar, A. Sagheer, Audio Visual Arabic Speech (AVAS) Database for Human-Computer Interaction Applications. The International Journal of Advanced Research in Computer Science and Software Engineering, **3**, 9 (2013).

Alaa Sagheer received his B.Sc. and M.Sc. degrees in Mathematics from Aswan University, Egypt. He received his Ph.D. in Computer Engineering (Intelligent Systems) from the Graduate School of Information Science and Electrical Engineering, Kyushu University, Japan in 2007. Currently, Sagheer is working as an Associate Professor at Aswan University, Egypt. However, since 2014 he joined the Department of Computer Science, College of Computer Science and Information Technology, King Faisal University, Saudi Arabia. Since 2010, Dr. Sagheer has been the founder, director and principal investigator at the Center for Artificial Intelligence and Robotics (CAIRO) at Aswan University. Recently, in 2013, Sagheer and his team won the first prize, in a programming competition organized by the Ministry of Communication and Information Technology (MCIT) Egypt, for his system entitled ”Mute and Hearing Impaired Education via an Intelligent Lip Reading System”. Sagheer’s research interests include pattern recognition, artificial intelligence, machine learning, human?computer interaction, and image processing. Recently, Sagheer extended his interest with quantum information, quantum communication, and quantum computer. Dr. Sagheer is a senior member of IEEE and a member of IEEE Computational Intelligence society.