Journal of Advances in Computer Research
Quarterly pISSN: 2345-606x eISSN: 2345-6078
Sari Branch, Islamic Azad University, Sari, I.R.Iran
(Vol. 8, No. 2, May 2017), Pages: 103-124
www.jacr.iausari.ac.ir

An Efficient Model for Lip-reading in Persian Language Based on Visual Word and Fast Furrier Transform Combined with Neural Network

Khadijeh Mirzaei Talarposhti✉, Mehrzad Khaki Jamei
Department of Computer Engineering, Sari Branch, Islamic Azad University, Sari, Iran
mirzaei.iausari@gmail.com; khaki.mehrzad@gmail.com

Received: 2016/10/07; Accepted: 2017/01/13

Abstract

Automatic lip-reading plays an important role in human computer interaction in noisy environments where audio speech recognition may be difficult. However, similar to speech recognition, lip-reading systems also face several challenges due to variances in the inputs, such as with facial features, skin colors, speaking speeds, and intensities. In this study a new method has been proposed for extracting features from a video containing a certain Persian words without any audio signal. The method is based on the fast furrier transform combined with the color specification of the frames in the recorded video of the spoken word. To improve the system performance visual word has been used as the shortest element of visual speech. Five speaker, three men and two women, have participated for capturing the videos of the spoken words. After obtaining features from the videos an artificial neural network has been employed as classifier. The experimental results show the average accuracy about 86.8% in recognition 31 Persian words.

Keywords: *Automatic Lip-Reading, Fast Furrier Transform, Visual Word, Artificial Neural Network*

1. Introduction

Automatic lip reading, also referred to as visual speech recognition (VSR) or sometimes speech reading, has received a great deal of attention in the last decade for its potential use in applications such as HCI, audio-visual automatic speech recognition (AV-ASR), speaker recognition, talking heads, sign language recognition and video surveillance [1]. A typical VSR system includes image acquisition, lip localization, feature extraction and recognition. The lips and the mouth region of a face reveal most of the relevant visual speech information for a VSR system. Therefore, it is important for any VSR system to focus on the lips area. While some approaches aim to directly locate the lips of the subject in question [2], others focus on the relatively easier task of locating the face and then locating the lips based on prior knowledge, e.g. the work of Zhang et al [3]. The most successful approaches to automatic lip reading depend mainly on recognizing a visual speech unit called a "viseme" (the visual part of a phoneme) [4–6]. A viseme is the shortest visually recognizable part of speech [1].

A general framework for AV-ASR has been developed during the last years [7,8], but for a practical deployment the systems still lack robustness against non-ideal working conditions. The first studies on AV-ASR with realistic conditions applied directly the systems developed for ideal visual conditions [9,10], obtaining poor lip-reading

performance and failing to exploit the visual modality in the multi-modal systems. Recently a study on audio-visual speech recognition has considered the multi-pose lip-reading [11]. The method is inspired by pose-invariant face recognition and relies on linear regression to find an approximate mapping between images from different poses. Although many of studies have used the visual features of lips and mouth to enhance the performance of the speech recognition [4,7–10,12], the others have consider the purely visual speech recognition [13–20]. The latter gives rise to a Silent Speech interface, which is defined as a system "enabling speech communication to take place when an audible acoustic signal is unavailable" [21]. Silent Speech technology has a large number of applications: It allows persons with certain speech impairments to communicate, as well as enabling confidential and un-disturbing communication in public places [21]. Further uses of lip-reading have been proposed, e.g. automatic speech extraction from surveillance videos and its interpretation for forensic purposes [22]. Lip-reading has been augmented with ultrasound images of the tongue and vocal tract [23,24]. Furthermore, there are Silent Speech interfaces based on very different principles, like speech recognition from electromyography [25–28] or (electro-) magnetic articulography [29].

There are two scenarios in the studies on VSR systems. Some studies consider the viseme as the shortest element of the visual speech [4,5,16], while a few others have focused on whole word in the speech recognition [1]. In 2009 A. B. Hassanat in his PhD thesis has shown that: detecting whole word instead of "viseme" as the smallest element of the speech can enhance the VSR systems [1]. He called this element the visual word. There are several studies on the Persian VSR systems that are based on the "viseme" [5,13,30–32], but, there is no study on the Persian VSR - that consider the whole word as the shortest recognizable element.

In this study a new approach has been developed for visual speech in Persian language based on the visual word. The study is conducted in four steps including: enhancing the feature extraction with trying the numerous relations between the levels of the colors in image to find the best formula for localization the lips on the image. Treating the values of each feature in the frames of video as a time domain function. Reducing the amount of data and neglecting the speed of speech by transforming the function to the phase frequency domain with fast furrier transform. Finally, training the artificial neural network with the dataset and testing the performance.

This paper is organized as follows: section 2 gives an overview of a typical VSR system. The proposed method is represented in section 3. Simulation of the proposed method is explained in section 4. The obtained results and comparison with the other methods are drawn in section 5, and finally, section 6 concludes the study.

2. Architecture of a VSR

A typical VSR system that uses visemes, as shown in Figure 1, includes image/video acquisition, lip detection, feature extraction, visemes recognition followed by word recognition based on their visemes [1].

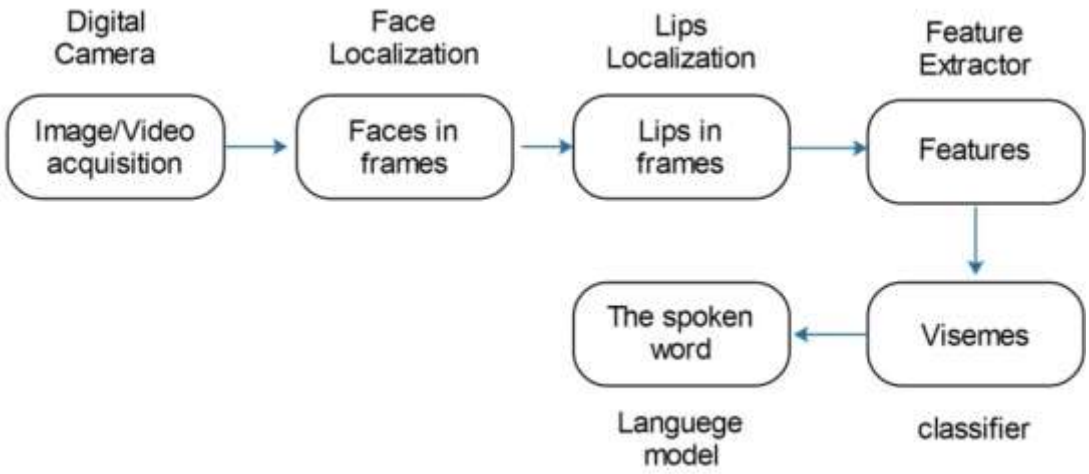
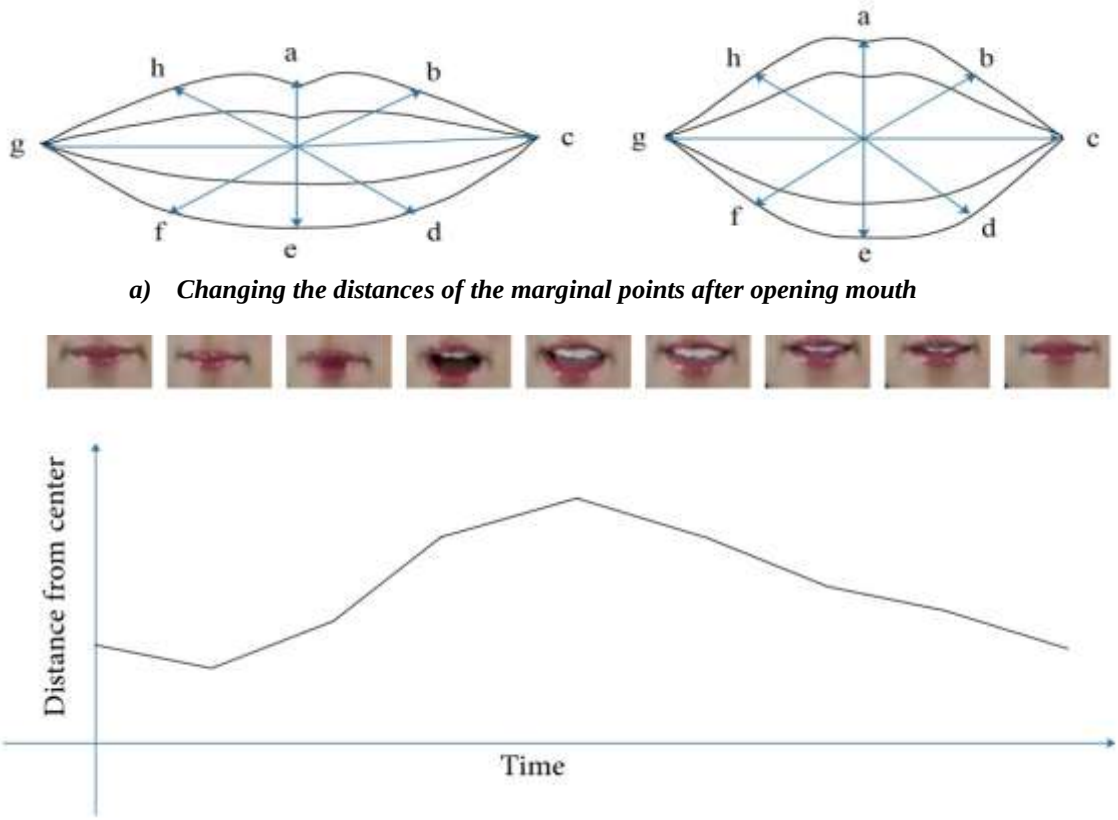


Figure 1. A typical viseme base VSR system **[1]**

During advances in VSR systems various algorithms have been used by researchers such as: Hidden Markov Model (HMM) [13,17], pattern matching [14], statistical [12], convolutional neural network (CNN) [33], unsupervised random forest [34], dynamic programming [35], support vector machine (SVM) [36], long short-term memory (LSTM) [22], local spatiotemporal descriptors [37], artificial neural network (ANN) [22], and clustering technics [5].

It is clear that for accurate lip-reading the extraction of features must be performed as well as possible. One of the most commonly used method that is used for extracting features of lips is called snake method [14]. In this method, as shown in figure 2-a, the distances of some points on the margins of the lips from center of the lips in the sequence of frames, make an snake like curve as shown in figure 2-b.



a) Changing the distances of the marginal points after opening mouth

b) The example of the snake curve for the point 'a' when the spoken word is 'mice'

Figure 2. Snake method for feature extraction

The lip-reading can be treated as a pattern recognition problem. Although, artificial neural networks (ANNs) have great potential in pattern recognition, there is a few works in the lip-reading that use ANN as the classifier. Suppose that, there are eight curves, each for one point of the lips margins, such as the curve shown in fig2-b. Hence, we have a large number of data that are treated as the input vector that must be trained with the classifier. Therefore, the amounts of data elements lead to the more complex ANN that many of researchers prefer to use the other classifiers as explained in the previous paragraphs.

3. Proposed Method

As previous discussed, using ANN as a classifier in lip-reading is depended on the number of elements of the input vector. Therefore, we use the fast furrier transform (FFT) of the curves for reducing the amounts of information contained in the curves without losing the main concepts of the curves. Furthermore, the use of FFT automatically neglects the speed of the spoken word, as will be explained in the next sections. In this section we explain the main contributions of the proposed method.

3.1 Using Visual Word Instead of Viseme

As previous discussed, the most studies in the lip-reading are based on recognize visemes as the shortest elements of the spoken word, and then with respect to the

recognized visemes another classifier must be used for recognize the spoken world by using the language model and the sequence of recognized visemes. In fact, viseme in the lip-reading is same as phoneme in the speech processing. Hence, a viseme is the shortest visually recognizable part of speech, and a phoneme is the smallest (shortest) audible component of speech. Typically, a phoneme is associated with a unique viseme or a sequence of visemes, but this is not true for all phonemes [38]. Hence, an automated VSR system that relies on visemes is faced with the difficult scenario of having to recognize words, which have some phonemes that share the same viseme(s) or phonemes that have no associated viseme.

Ahmad B. Hassanat in his PhD thesis in 2009 [1] showed that several problems arise while using visemes in visual speech recognition systems [1]. For example: some phonemes have different visual effects (visemes) in the various situations, such as 'n' in the words "banana" and Nottingham". Furthermore, certain phonemes have week visual effects, because they articulated from inside of the mouth [1]. According the above issues we decided to use visual word instead of viseme in the proposed VSR system.

3.2 Lip Localization

It is clear that for finding an object in the image it must has a differentiable property. The most common methods in the VSR systems have used the colors of images for lip localization. As we know, a colored image has three levels of colors: red, green, and blue. Several studies have used the proportions between red and the others to find the region of lips. For a clear see on the relation between colors, figures 3b-d show the several relations between the levels of colors for the pixels that are located on the column 160 of the example image that is stated with the blue line in figure 3-a.

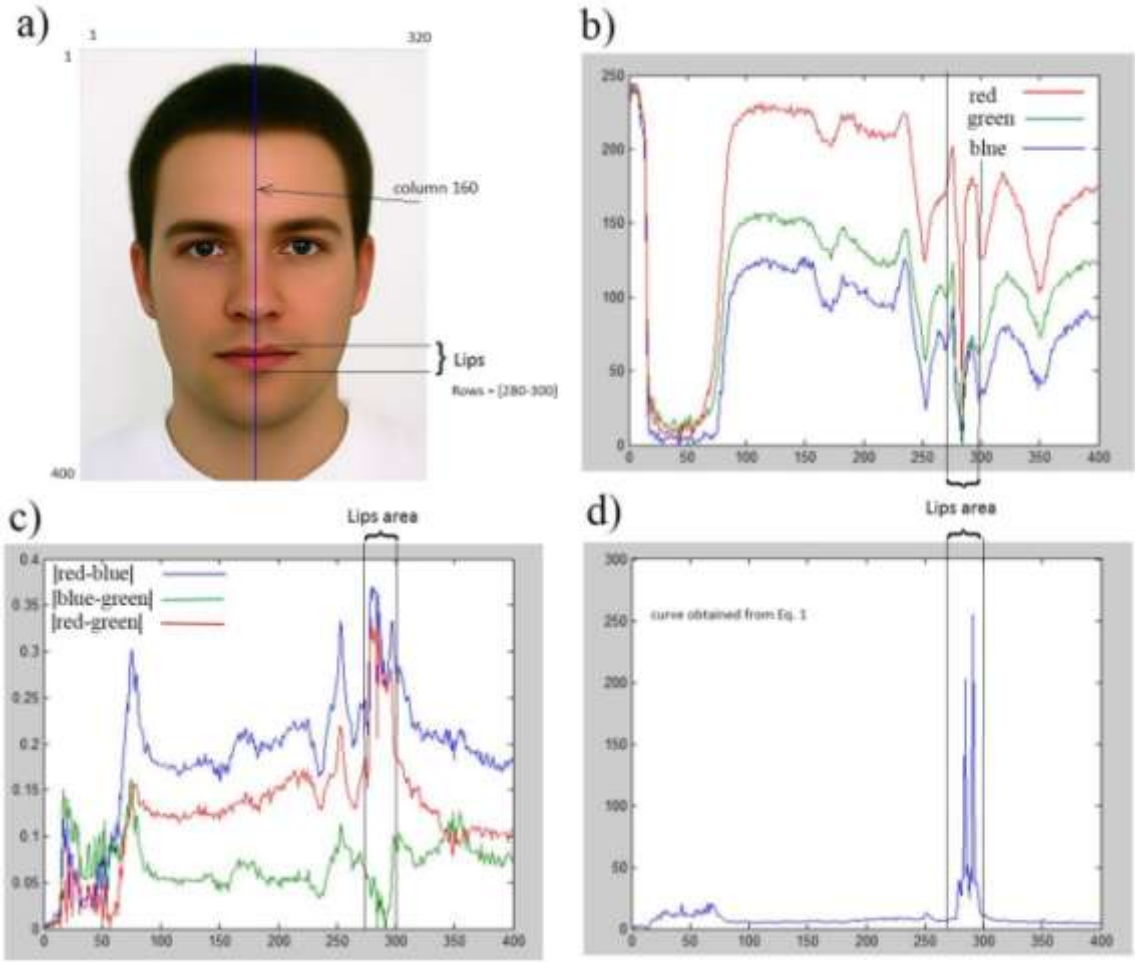


Figure 3. Relation between levels of colors in the example image.
a) Example image. b) Levels of the colors red, green, and blue for pixels of the stated column. c) The normalized absolute difference between colors. d) The values for the Eq. 1 for the levels of colors in the stated column

In figure 3-b, we can see the pure levels of colors for the pixels that are located on the stated column. As we can see, there is no distinguishable appearance between colors in figure 3-b. In figure 3-c, the lips area is more tractable, but, it is not clear enough. Finally, the best distinguishable formula is shown in figure 3-d, which is obtained by Eq. 1 as follows:

$$f = \frac{|red - blue| + |red - green|}{1 + |blue - green| \times (red + green + blue)} \quad (1)$$

After localization lips the remaining of image is eliminated and a rectangle with the minimum size that contains the entire lips would be consider in the rest of algorithm. The major steps for localization lips are shown in figure 4.

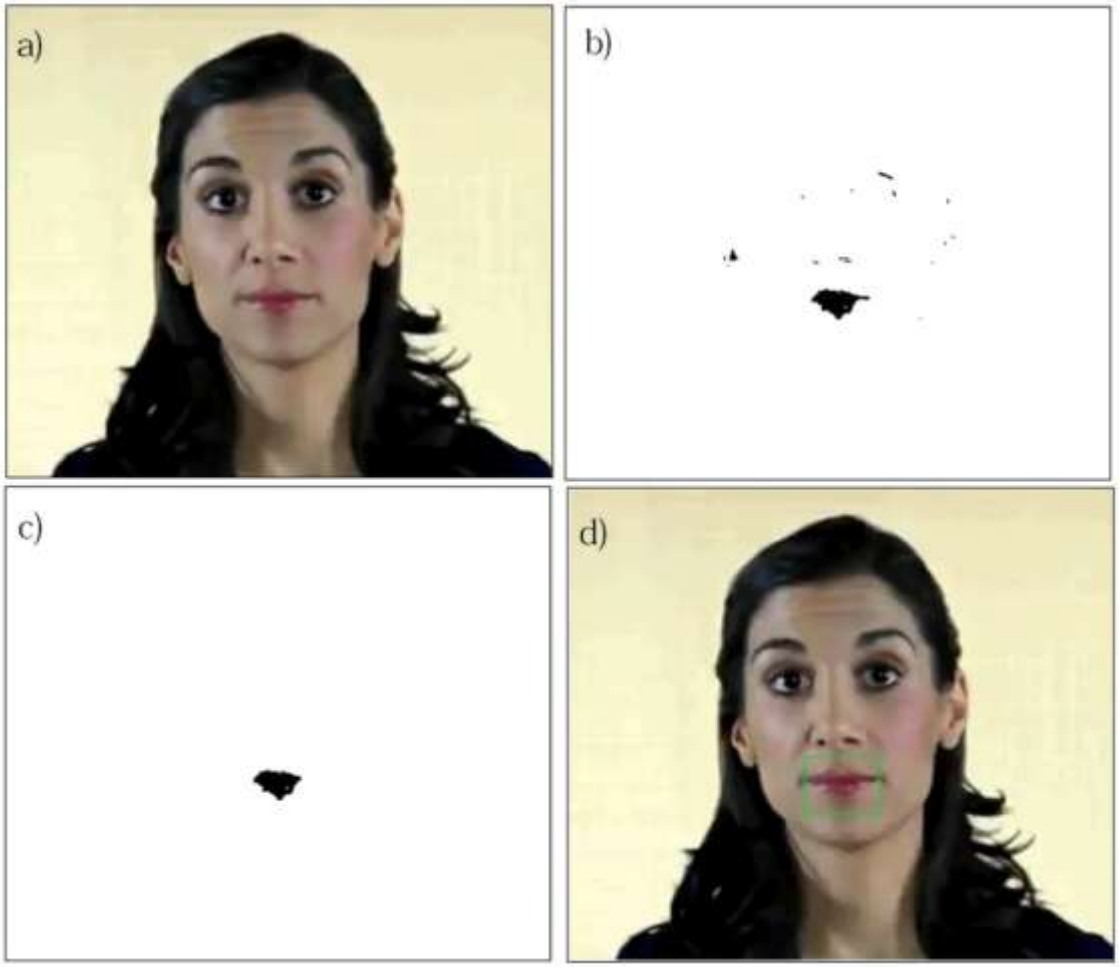


Figure 4. The major steps for localization lips.
a) Original image. b) The pixels that obtained from Eq. 1. c) The pixels of the lips after removing the noise pixels. d) The lips are located by the green rectangle on the face.

3.3 Feature Extraction

By trying many different features, finally, four features, which are shown in figure 5, have been used as follows: 1) Height-to-width ratio that represents the amount of opening the lips. 2) The Area of the lips that shows the thickness of lips. 3) The Area of teeth that represents depth of the lips. 4) The Area of the tongue that shows the role of tongue on the spoken word.

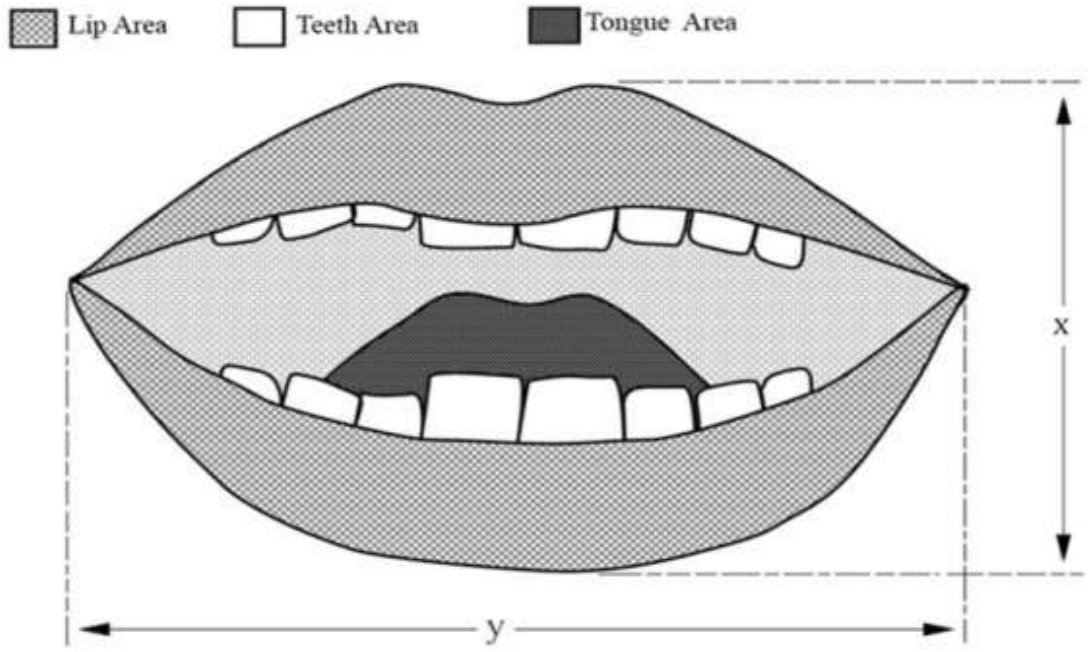


Figure 5. The chosen features of image: (x/y), lip area, teeth area, and tongue area

To calculate the features of each frame, after obtaining the minimum rectangle that contains the entire lips (figure 6-a), an algorithm that is shown in Table 1, makes a mask that could be used to remove all of the pixels that are outside of the lips.

Table 1. The algorithm for making a mask for removing the pixels out of the lips.

Inputs: *The minimum sized rectangle 'A' as a 3D matrix ($h^1 \times w^2 \times 3$) that contains the lips (figure 6-a);*
Output: *a Boolean matrix 'M' in size ($h \times w$) that is the mask for lips (1 for pixels inside lips & 0 for others);*

```
1    All elements of M_{h×w} := 1;
2    FOR i := 1 TO w DO
3      FOR j := 1 TO h DO
4        Pixel := A(i, j, 1-3);
5        IF Pixel has not the conditions of the lips' pixels THEN
6          M(i, j) := 0 ;
7        ELSE
8          BREAK;
9        END IF
10     END j
11   END i
12   FOR i :=1 TO w DO
13     FOR j := h DOWNTO 1 DO
14       Pixel := A(i, j, 1-3);
15       IF Pixel has not the conditions of the lips' pixels THEN
16         M(i, j) := 0;
17       ELSE
18         BREAK;
19       END IF
20     END j
21   END i
```

[1] *h: the height of the rectangle that is equal to the height of lips;*
[2] *w: the width of rectangle that is equal to the width of lips;*

An example of the mask generated by the algorithm of Table 1 is sown in figure 6. In the figure 6, an area of interest, which contains the entire lips, could be seen in figure 6-a. this area obtains from a frame of a spoken word. The outcome of the algorithm is shown in figure 6-b.

Figure 6. An example of masking the lips.

a) Original region of lips located by proposed algorithm. b) The mask of the lips generated from (a) by the algorithm of Table 1. c) Applying the mask (b) on the lips image (a)

It is clear that after filtering the frame of the spoken word to the masked lips as shown in figure 6-c, detection the other areas such as teeth, tongue, and lips becomes more simple than the initial frame. Therefore, another algorithm simply separates these areas (Table 2). The algorithm gets the mask (figure 6-b) and the image of the lips (figure 6-a) as the input and generates a matrix same size as the lips' image region (figure 6-a). The matrix contains integer numbers that each element delegates the pixel of the lips image in its corresponding location. The matrix divides the image pixels into four layers including: lips, teeth, tongue, and inside. Indeed, the number of elements in each layer delegates the amount of area in that layer in the lips' image.

Table 2. The algorithm for separating the layers of the lips' image.

Inputs: *The lips' image 'A' as a 3D matrix ($h^1 \times w^2 \times 3$) & the mask of the lips 'M' ($h \times w$) generated from Table 1.*

Output: *Matrix 'L' ($h \times w$) of type integer that states[3] the layers of the pixels in the lips' image.*

```
1    FOR i := 1 TO w DO
2      FOR j := 1 TO h DO
3        IF M(i, j) = 0 THEN
4          L(i, j) := 0;
5        ELSE
6          Pixel := A(i, j, 1-3);
7          IF Pixel has the conditions of the teeth's pixels THEN
8            M(i, j) :=1 ;
9          ELSE IF Pixel has the conditions of the lips' pixels THEN
10           M(i, j) :=2 ;
11         ELSE IF Pixel has the conditions of the cavity's pixels THEN
12           M(i, j) :=3 ;
13         ELSE
14           M(i, j) :=4 ;
15         END IF
16       END IF
17     END j
18   END i
```

[1] *h: the height of the lips' image;* [2] *w: the width of the lips' image;* [3] *the integer number that is contained with matrix L, shows the layer of corresponding pixel of lips' image as follows: 0 → outside of lips , 1 → teeth, 2 → lips, 3 → cavity of mouth, 4 → tongue;*

The algorithm seen in Table 2 was designed such a way that the layers recognized from simple to difficult. It is clear that by removing the simpler layer, because of reduction the area, pixels, and complexity, the next layer becomes more recognizable. Figure 7 shows the steps of the algorithm.

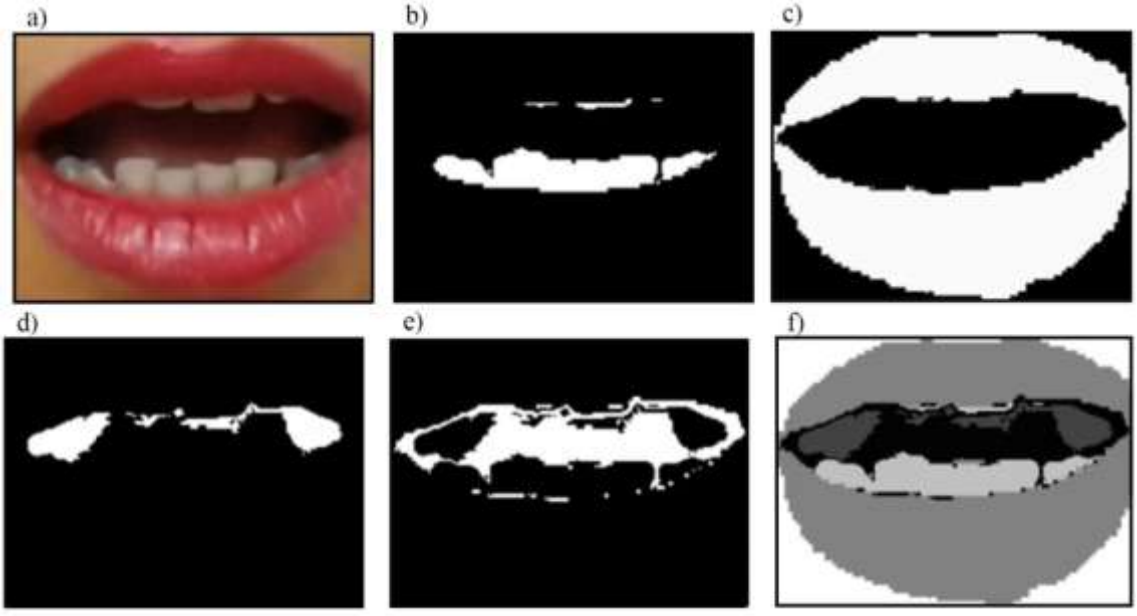


Figure 7. The steps of the algorithm.

a) The original image of the lips. b) Teeth recognized by the proposed method. c) In the next step the pixels of lips are marked. d) After the lips' pixels the mouth cavity is recognized. e) The remaining pixels make up the tongue. f) All layers are shown in one image, each layer with a specific gray level.

Currently, for each frame of the spoken word, we can measure the area of each layer by counting the number of pixels in each layer. In the next section we show the processing of the obtained values of the layers.

3.4 Signal Processing

As previous discussed, one of the most popular methods for processing the obtained signals in the lip-reading is called: snake. In this study a similar method has been used for signal processing. There are four features extracted from each frame of the spoken word. These features that were explained in the previous section, for a sample spoken Persian word 'salaam' (means hello) has been shown in figure 8. The frames were obtained from a MPEG4 video with the frame rate about 30 frames per second. The video duration was about two second that contains approximate 60 frames. Figure 8 shows the layers and features for 10 frames of the video with the equal interval (interval= 6 frames).

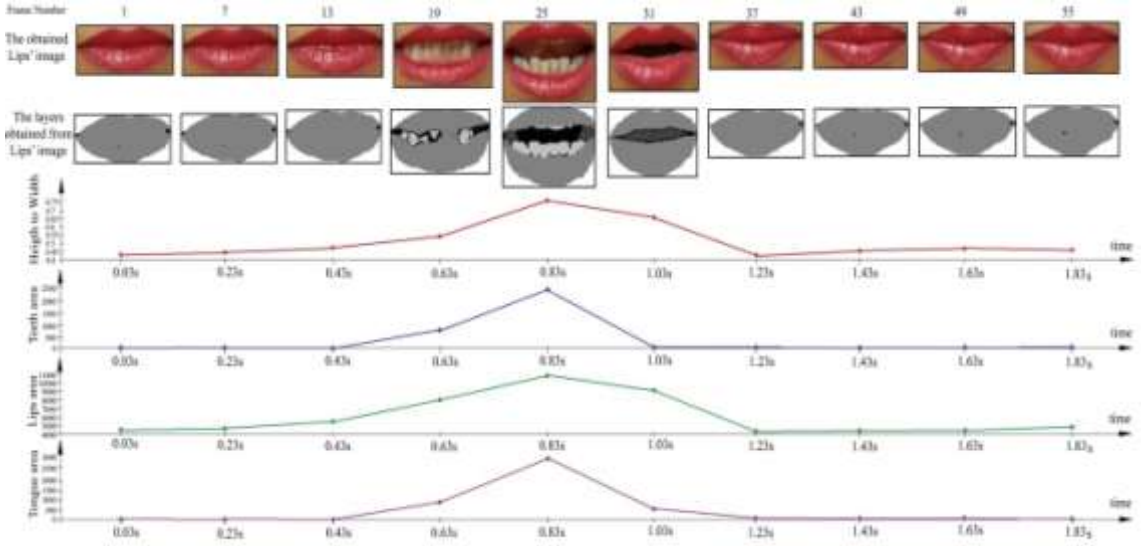


Figure 8. Signals obtained from the features of the frames of the Persian word 'Salaam'

3.5 Removing the Frames Before and After the Spoken Word

As can be observed in figure 8, even in a short interval of time (for example two seconds in figure 8) there are many frames that have not significant information to be used in speech recognition. Hence, these redundant frames must be removed from dataset. With a clear see on the figure 8, we can simply find that the frames (1-13 & 37-60) are approximately same. Because, the lips in these frames are closed and this situation is not depended on the spoken word. In the other word, these frames are related to the situations that the spoken word has not been started, or has just been finished.

Currently, we must design an algorithm for detection the frames confined to the spoken word. Notice to the figure 8: total area of the lips including tongue, teeth, and cavity can be used in the algorithm. This value can be simply obtained from the mask of the frame, which is obtained by the algorithm in Table 1. The outcome of the algorithm that is shown in figure 6-b specifies the area mentioned above. It is clear that, if the change on the area is less than a threshold for a specific time, it means that the frames are out of the spoken word's confine. We use Eq. 2 to calculate the criterion for detection the frames related to the spoken word.

$$\begin{cases} y(\mathrm{t}) = \beta(t) \times \left| f'(t) \right| \\ \beta(t) = \dfrac{f(t) - \min(f(t))}{\max(f(t)) - \min(f(t))} \end{cases} \tag{2}$$

Where t denotes the frame number or time, and $f(t)$ represents the lips' area for the frame that is specified by t and $f'(t)$ is derivation of $f(t)$. To explain the above Equation, notice the derivation of a function shows the rate of changes on the values of function. In the other hand, when the lips' area is in its minimum value it means that the lips are closed. Therefore, we used the multiply of the normalized form of the lips' area (denoted by $\beta(t)$ in Eq. (2)) and its derivation to detect the spoken word's confines. Additionally, because, the sign of change's rate is not important in this manner, we used

the absolute values of $f'(t)$ in the product. An example is shown in figure 9 that the values of the lips' area $\left(f(t)\right)$, its derivation $\left(f'(t)\right)$, absolute of derivation $\left(\left|f'(t)\right|\right)$, and the final product $\left(y(t)\right)$ for a Persian spoken word 'khubam' (means I'm fine) is observable.

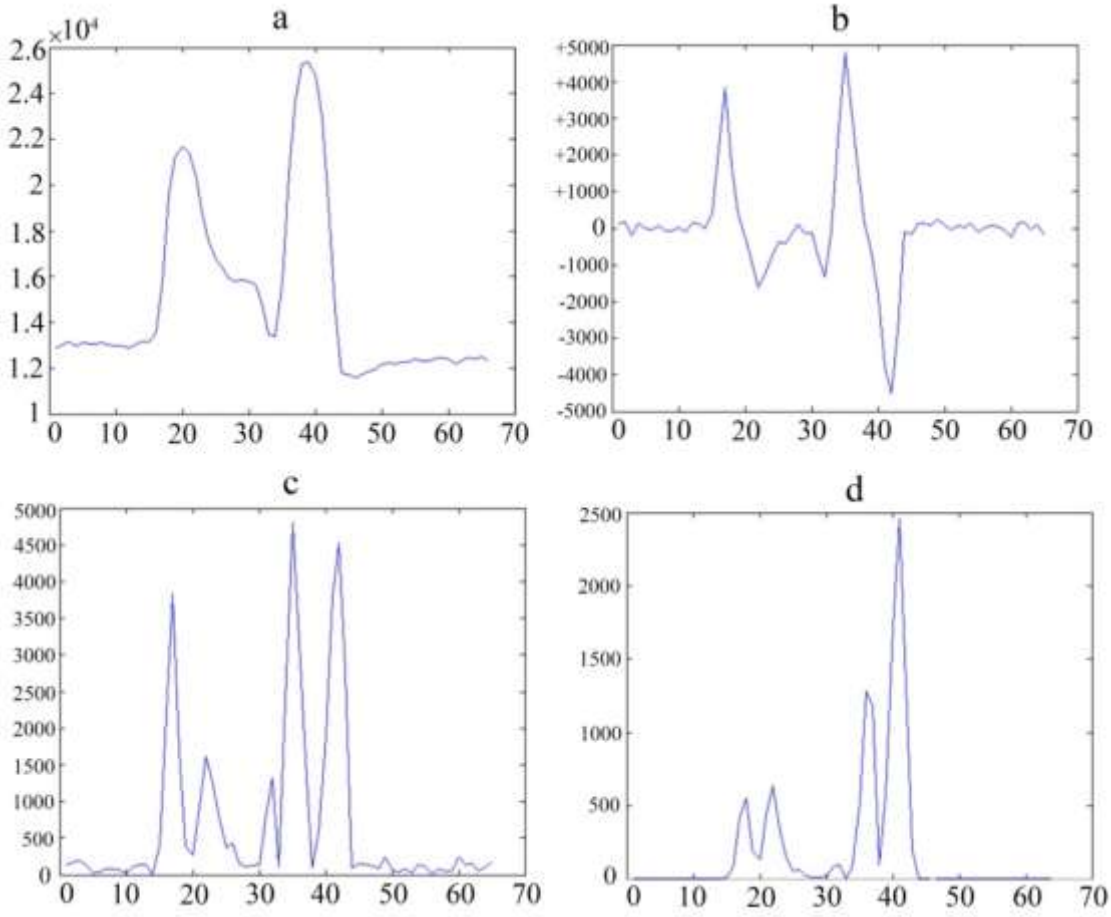


Figure 9. Removing the frames out of the spoken word's confines.
a) The signal of the lips' area for the Persian word 'khubam'. b) Derivation of the (a). c) The absolute values of (b). d) Multiply for normalized (a), (b), and (c).

As seen in figure 9, the confine of the sample spoken word is throughout frames (15-45). Notice the figures 9a-d, the best figure that can separate the spoken word's frames with the others, is figure 9-d, which is obtained from Eq. 2. The obtained values for the function of figure 9-d, in the frame numbers < 15, or > 45 are approximately equal to zero. In this study, we choose 0.005 as a threshold for detection redundant frames. The value for threshold has been obtained by trial and error.

3.6 Applying Fast Fourier Transform on the Signals

The final step of the proposed VSR system employs artificial neural network (ANN) as the classifier of the collected signals. As previous discussed, there are four signals extracted from the video of the spoken word. The signals are obtained from the features of the video of the spoken word including: lips' area, teeth area, tongue area, and the proportion of lips' height to the lips' weight. For the sake of reduction the number of

inputs in the proposed neural network, we used the furrier coefficients of the signals as the ANN's input. As we know the furrier series could be shown by Eq. 3 as follows:

$$f(t) = \frac{a_0}{2} + \sum_{n=1}^{\infty} \left[a_n \cos(\omega_n t) + b_n \sin(\omega_n t) \right] \quad (3)$$

Where a_n and b_n are the furrier coefficients. To calculate the furrier coefficients of a vector, we can use Eqs. (4-6), which is called fast furrier transform (FFT). The FFT, also referred to Discrete Furrier Transform (DFT), converts a signal from its original domain (often time or space) to frequency domain.

$$a_0 = \frac{1}{2\pi} \sum_{i=1}^{n} f(i) \quad (4)$$

$$a_n = \frac{1}{\pi} \sum_{i=1}^{n} f(i) \cos(nx) \quad , \quad x = 2\pi \left(\frac{i}{n} \right) - \pi \quad (5)$$

$$b_n = \frac{1}{\pi} \sum_{i=1}^{n} f(i) \sin(nx) \quad , \quad x = 2\pi \left(\frac{i}{n} \right) - \pi \quad (6)$$

Although, there are extreme coefficients in the furrier equation (Eq. 3), practically, only a few coefficients are necessary to have a good approximation of the original function. Indeed, the coefficients with high index have not a significant effect on transformation. For example, in figure 10, a signal of the lips' area of the Persian spoken word 'khubam' (figure 9-a) after removing redundant frames by the algorithm explained in 3.5, (figure 10-a) has been transformed by the FFT algorithm with different numbers of coefficients (figure 10b-i).

As can be observed in figure 10b-i, when the number of coefficients goes high, the function reconstructed from the furrier coefficients becomes more likely the original signal (figure 10a). By the way, when the number of coefficients is large enough, increasing the number of coefficients doesn't have a significant change on the generated function (see figure 10-h and i).

In this study, we choose (n=4) for transforming the signals of the features. As explained in previous sections there are four features that are represented as the signals in the time domain. Hence, with (n=4) after transforming the signals, there must be 9 real numbers for each signal that are corresponding the coefficients (a_0,…,a_4) and (b_1,…,b_4). For above discussion, finally, there are 36 real numbers as a vector for the input pattern of the neural network that is designed for recognizing the spoken word.

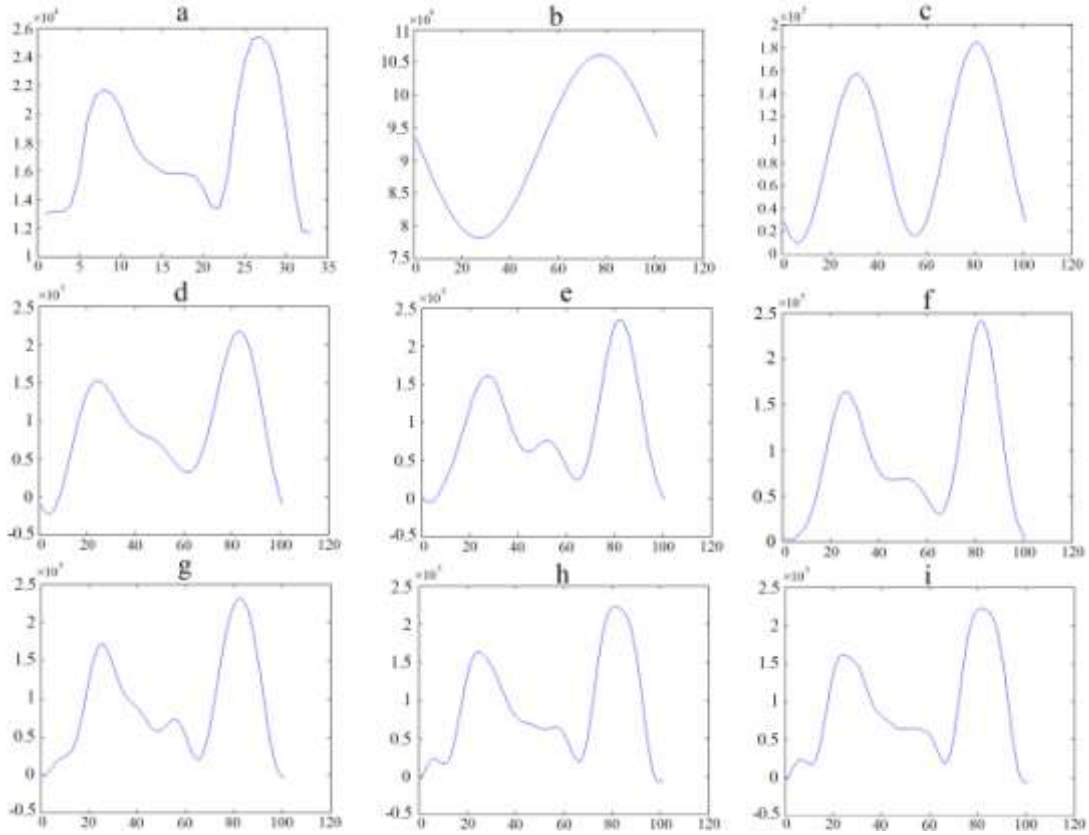


Figure 10. The fast furrier transform of the signal of the Persian word 'khubam' Original signal (a) has been transformed with n=1 (b), n=2 (c), n=2 (d), n=4 (e), n=5 (f), n=6 (g), n=8 (h), and n=10 (i).

3.7 Artificial Neural Network

As explained in previous section, there are 36 real numbers as the inputs of classifier. Lip-reading can be formulated as a pattern recognition problem. Artificial neural networks (AANs) have great potential in the pattern recognition [39–42]. The ANN that is chosen in this study in called Multi-Layer Perceptron (MLP) that consists of three types of layers of artificial neurons including: input layer, hidden layer(s), and output layer. The number of hidden layers is depended on the complexity of the problem. A popular method to determine the number of hidden layers is trial and error. In this study an MLP with two hidden layers has been employed for recognition certain Persian spoken words. The employed MLP is shown in figure 11. Typically, an artificial neuron can be determined by Eq. (7) in which an activation function converts the input to the desired output.

$$a = f(N + b) \tag{7}$$

Where a is the output and b is the bias of the neuron. N is weighted input that can be obtained from Eq. (8) as follows:

$$N = \sum_{i=1}^{m} w_i p_i \tag{8}$$

Where m is the number of inputs, p_i is i'th input, and w_i is the weight of the i'th input of the neuron. The most common activation functions are: linier, sign, sigmoid, and hyperbolic tangent (signed sigmoid). In general, a layer of ANN is represented in the matrix form. Suppose that in a layer, there are n neurons that each one has m inputs. In this form the weights of neurons are treated as a matrix with n rows and m columns, and the input and bias of the neurons could be vectors with the size of $(m\times1)$ and $(n\times1)$, respectively. For this layer Eq. (9) shows the output of the layer, which is a $(n\times1)$ vector. Indeed, a layer of neural network with m neurons converts the input vector from n dimensional space to the m dimensional space.

$$\begin{cases}\vec{a}_{n\times1}=f\left(\vec{N}_{n\times1}+\vec{b}_{n\times1}\right)\\ \vec{N}_{n\times1}=\vec{W}_{n\times m}\times\vec{P}_{m\times1}\end{cases} \tag{9}$$

Where $\vec{p}$, $\vec{W}$, $\vec{b}$, and $\vec{a}$ are the input vector, weights matrix, biases, and the output vector, respectively. The employed neural network for this study is shown in figure 11. As shown in figure 11, the neural network has two hidden layers. The number of neurons in the input layer, first and second hidden layers, and the output layer are 36, 16, 20, and 5, respectively. Hence, the final output of the ANN must be a five dimensional vector.

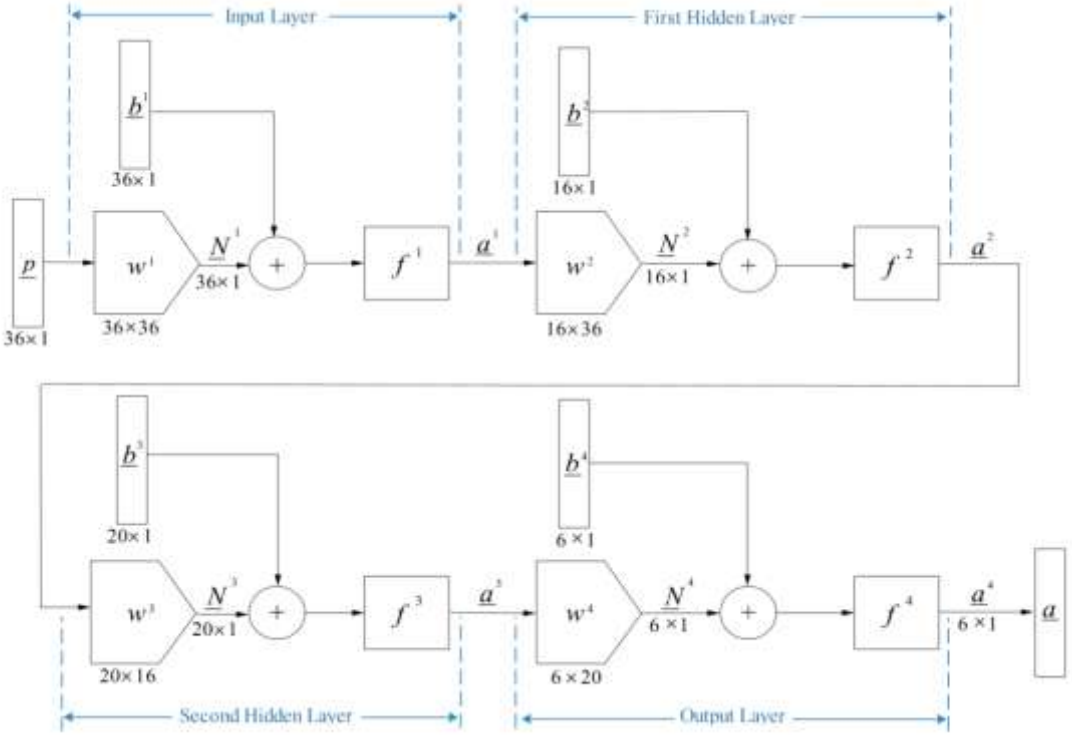


Figure 11. The ANN that is used in this study.

4. Simulation the Proposed Method

The explained neural network is simulated in MATLAB's neural network toolbox. As discussed in previous section we have four series of features that extracted from a video containing a Persian spoken word. Each signal after furrier transform converts to nine real number that are the furrier coefficients for the signal. We have totally 36 furrier coefficients for the signals. Hence, the input pattern is a 36 dimensional vector that must be apply to the neural network. Figure 12 briefly shows the architecture of the proposed VSR system.

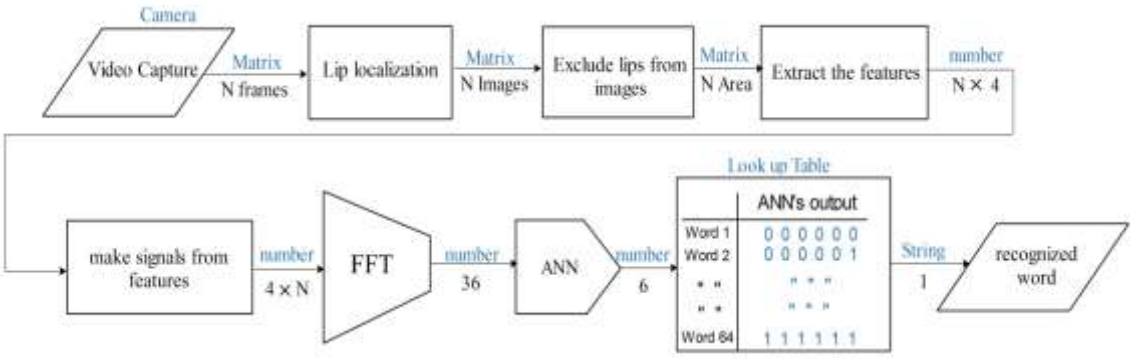


Figure 12. Architecture of the proposed VSR.

4.1 Dataset

In this study 31 commonly used Persian words have been consider for recognizing by the proposed VSR system. Additionally, several Persian words were added to the dataset for recognition unknown words as the neural network generates the output="others". In the other words, ANN generates the word's Id =31 that means the word is not in the set of the known words. To achieve that, each word is spoken at least 10 times in different conditions of light and distance, by three men and two women. Hence, we have at least 1550 different videos of the known spoken words and 500 video of the unknown spoken words. However, we have totally 2050 videos of the known and unknown Persian words. The videos are in the JPEG4 format and the parameters of videos are 29.86 frames per second with the resolution 1920×1080 pixels that the colors standard is RGB24.

4.2 Training ANN

The employed ANN (figure 11) is a multi-layer perceptron (MLP) with four layers. The numbers of neurons in the Input, first and second hidden and output layers are 36, 16, 20, and 5, respectively. Activation functions for input and output layers are linier and sigmoid, respectively, and for the hidden layers are hyperbolic tangent or signed sigmoid.

Implementing the ANN is performed in the MATLAB's neural network toolbox. The training algorithm is Levenberg–Marquardt back propagation (TRAINLM). Back propagation is the most popular training algorithm that was introduced by Pol J. Werbos in his PhD thesis in 1974 [43]. Furthermore, the chosen learning function was gradient descent with momentum (LEARNGDM) and the error function was sum square errors (SSE), which could be obtained by Eq. (10) as follows.

$$SSE = \sum_{i=1}^{n}\sum_{j=1}^{m}\left(a_j^{(i)} - d_j^{(i)}\right)^2 \quad (10)$$

Where n and m are the numbers of training patterns and the network's outputs, respectively. $a_j^{(i)}$ denotes the $j'th$ output of the network for $i'th$ training pattern, and $d_j^{(i)}$ represents its corresponding desired output.

5. Results and Comparisons

In this section, first the proposed method is evaluated and some discussion on the obtained results will be done. Then the results will be compared with the certain existing methods in literature.

5.1 Evaluation the Proposed VSR System

Although, Cross Validation is the most popular validation algorithm for evaluation models, we used Hold Out validation algorithm for testing the performance of the proposed method for the sake of the following reasons. Choosing the validation algorithm is depended on the certain aspects of problem and the amount of dataset. Cross Validation is the best choice when there are a few numbers of samples. For example, when collecting samples are so difficult or expensive. When the amount of samples is large enough, Hold Out gives the most reliable results in the simulation of the model.

As explained in previous section, there are at least 2050 samples of the 31 Persian spoken words and certain unknown Persian words. Therefore, we used Hold Out validation method, because the number of collected samples is large enough. To employ Hold out system, samples are divided into the training, and testing samples with 850 and 1200 samples respectively. Dividing samples is performed randomly with a condition that the training dataset must contains all of the 31 spoken words and the enough number of unknown words. These words are shown in Table 3.

Table 3. Certain Persian words that are to be recognized by the proposed method.

Id	Word	meaning	Id	word	meaning	Id	word	meaning	Id	word	meaning
0	Sæla□m	Hello	8	færda□	Tomorrow	16	di□va□r	Wall	24	sib	Apple
1	khu□bæm	I'm fine	9	emru□z	Today	17	mi□z	Table	25	kela□s	Class
2	tʃ etori□	How R U?	10	diʃ æb	Last night	18	sændæli□	Chair	26	dærs	Curse
3	mæmnu□n	Thank U	11	emʃ æb	Tonight	19	dæftær	Office	27	sæfær	Trip
4	keta□b	Book	12	hæfteh	Week	20	Ta□xi	Taxi	28	na□n	Bread
5	dær	Door	13	ma□h	Moon	21	ma□shi□n	Car	29	gu□ʃ	Ear
6	ʃ æb	Night	14	sa□l	Year	22	dæst	Hand	30	tʃ eʃ m	Eye
7	ru□z	day	15	Ghærn	Century	23	pa□	Foot	31	------	Other

The obtained result from testing the proposed method are shown in figure 13. The results show that the proposed VSR system has a good performance in recognition Persian words. Additionally, considering unknown words in the training process may cause the significant reduction in system's performance. This issue was neglected by the existing studies in literature.

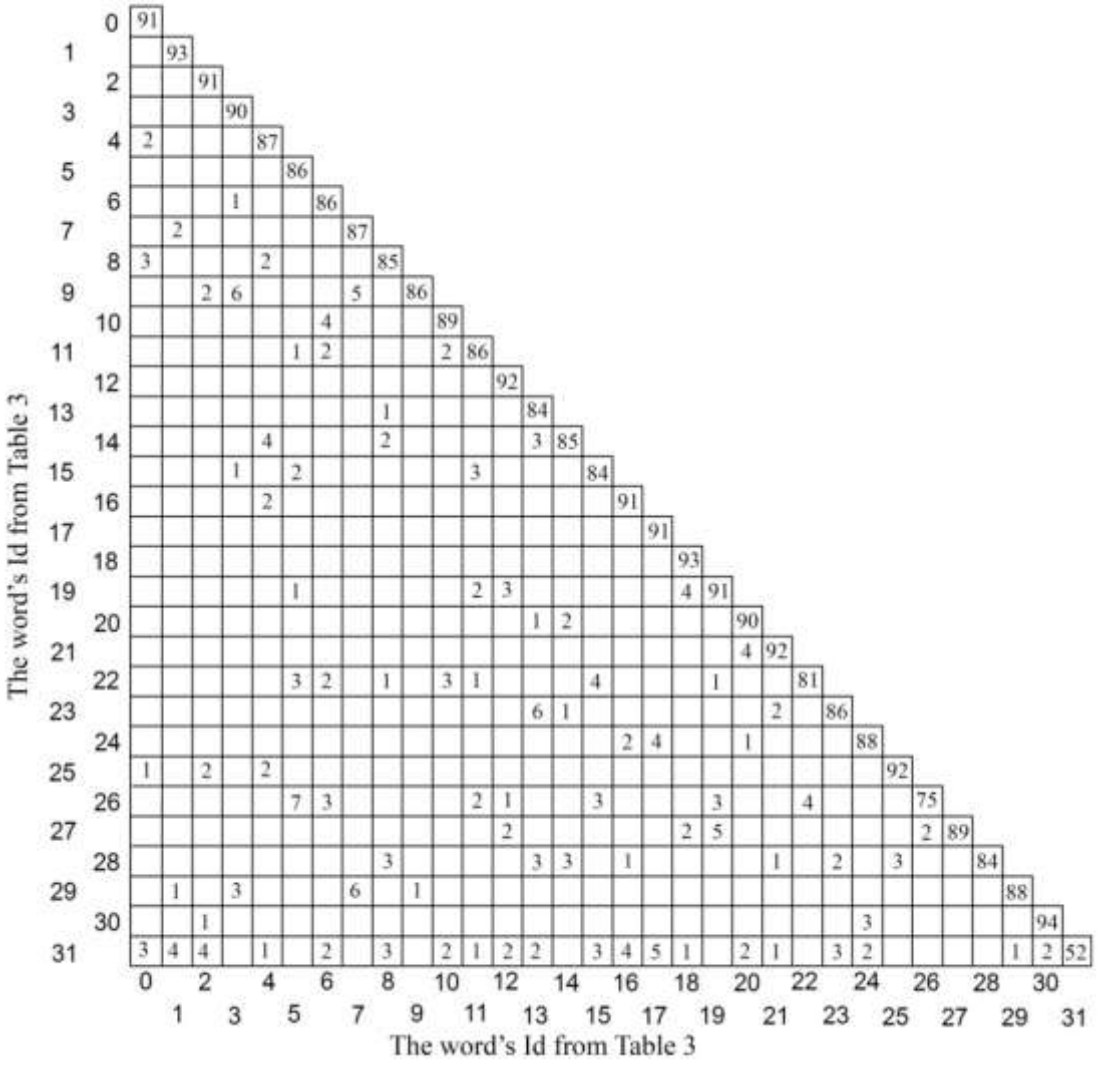


Figure 13. The results obtained from testing the ANN (numbers in grids denote percent)

As can be seen in Figure 13, some words are less confused with the others. For example the Persian words "khu□bæm" and "sændæli□" that are indexed by 1 and 18, respectively, with the accuracy about 93% are more recognizable than the words such as "dærs". With the similar analysis as the viseme in section 3.1 when the word's duration is large enough, the word would be more recognizable, and for the shorter words the system would be more confused. Naturally, the short words behavior is similar to visemes. The average accuracy for recognition the chosen Persian words including certain unknown words is 86.8%.

5.2 Comparisons with the other Works

Although comparison between two VSR systems must be performed in the same languages, for the sake of the lack of works in Persian language we have to compare the proposed VSR system with the both Persian and non-Persian languages. Additionally, we compare the viseme-based and word-based methods with the proposed method in terms of accuracy. Table 4 shows the comparisons in details.

5.3 Discussion

As can be observed in Table 4, the proposed method has higher accuracy than the other methods. But, what is the reason for such a high accuracy in testing the proposed method. Although, throughout previous sections detailed explanations have been represented, a brief discussion explains the obtained performance of the proposed VSR system as follows. From Table 4 and discussions on the section 3.1, it is clear that generally the systems that use 'word' as the shortest recognition element are more accurate than the ones that use 'viseme'. In the other hands, an artificial neural network with a suitable design and adequate number of layers and neurons is an extreme good performance in the pattern recognition problems. But, in this problem the amount of data significantly increases complexity of the ANN. Hence, we reduced the amount of data elements by applying FFT on the input signals. Therefore, a vector with 36 elements as the MLP's input is quite suitable to training the employed four layers MLP (figure 11).

Table 4. comparisons between the proposed method and the other works

Author	Year	Method	Language	Recognition element	Number of words to be recognized	Accuracy
T. shinchi et. al. [44]	1998	ANN[1]	Japanese	viseme	5	70%
W.S. Sadeghi & K. Yaghmaei [20]	2006	K-means + ANN	Persian	viseme	6	64.4%
M. Wand et. al. [22]	2016	Eigen lips + SVM[2]	English	word	26	70.6%
M. Wand et. al. [22]	2016	HOG[3] + SVM	English	word	26	71.3%
M. Wand et. al. [22]	2016	LSTM[4]	English	word	26	79.6%
Proposed method		**FFT[5] + MLP[6]**	**Persian**	**word**	**31**	**86.8%**

[1] Artificial Neural Network; [2] Support Vector Machin; [3] Histograms of Oriented Gradients; [4] Long Short-Term Memory; [5] Fast Furrier Transform; [6] Multi-Layer Perceptron;

6. Conclusion

In this study, a novel method for extraction features from a video containing a Persian spoken word, for the aims of lip-reading system, has been proposed. The method is based on color specification of the frames combined with fast furrier transform (FFT). Additionally, to obtain higher performance the visual effects of the word articulation has been considered as the shortest element of speech. This element, which is called visual word, is used in the proposed system instead of viseme. The extracted features are used as the input of a multi-layer perceptron (MLP) as the classifier. The number of Persian words, which are spoken by five speakers (three men and two women), are 31. The experimental results revealed that the accuracy of the proposed method is comparable with the other state of the arts in the visual speech recognition (VSR) systems.

References

[1] A.B. Hassanat, "Visual Words for Automatic Lip-reading," PhD Thesis, university of Buckingham, 2009.

[2] E. Gomez, C. Travieso, J. Briceno, M. Ferrer, "Biometric identification system by lip shape," 36th Annu. Int. Carnahan Conf. Secur. Technol., pp. 39–42,2002.

[3] J.M. Zhang, L.M. Wang, D.J. Niu, Y.Z. Zhan, "Research and implementation of a real time approach to lip detection in video sequences," Int. Conf. Mach. Learn. Cybern., pp. 2795–9,2003.

[4] J. Melenchon, J. Simo, G. Cobo, E. Martinez, "Objective viseme extraction and audiovisual uncertainty: estimation limits between auditory and visual modes," Int. Conf. Audit. Speech Process, 2007.

[5] M. Aghaahmadi, M.M. Dehshibi, A. Bastanfard, M. Fazlali, "Clustering Persian viseme using phoneme subspace for developing visual speech application," Multimed Tools Appl, Vol. 65, pp.521-41, doi:10.1007/s11042-012-1128-7, 2013.

[6] S. Werda, W. Mahdi, A. Ben-Hamadou, "Lip Localization and Viseme Classification for Visual Speech Recognition," Int J Comput Inf Sci, Vol. 15, 2007.

[7] G. Potamianos, C. Neti, G. Gravier, A. Garg, A. Senior, "Recent advances in the automatic recognition of audio-visual speech," Proc IEEE 2003, Vol. 91, pp.1306–26, 2003.

[8] G. Potamianos, C. Neti, J. Luettin, I. Matthews, "Audio-visual automatic speech recognition: an overview," In: G. Bailly, E. Vatikiotis-Bateson, P. Perrier, editors, Issues Audio-v. speech Process., Cambridge, MA: MIT Press; 2004.

[9] K. Livescu, O. Cetin, M. Hasegawa-Johnson, S. King, C. Bartels, N. Borges, et al., "Articulatory feature-based methods for acoustic and audio-visual speech recognition," 2006.

[10] G. Potamianos, C. Neti, "Audio-visual speech recognition in challenging environments," Eighth Eur. Conf. Speech Commun. Technol., p p. 1293–6, 2003.

[11] V. Estellers, J. Thiran, "Multi-pose lipreading and audio-visual speech recognition," EURASIP J Adv Signal Process 2012, Vol. 51, pp.1–23, 2012.

[12] S. Nakamura, "Statistical Multimodal Integration for Audio–Visual Speech Processing," IEEE Trans Neural Networks, Vol. 13, 2002.

[13] S. Alizadeh, R. Boostani, V. Asadpour, "Lip Feature Extraction and Reduction for HMM-Based Visual Speech Recognition Systems," 9th Int. Conf. Signal Process., pp. 561–4, 2008.

[14] M. Barnard, E.J. Holden, R. Owens, "Lip tracking using pattern matching snakes," 5th Conf. Comput. Vis., pp. 1–6. 2002.

[15] G.I. Chiou, J.N. Hwang, Lipreading from Color Video," IEEE Trans Image Process, Vol. 6, pp.1192–5, 1997.

[16] T. Ezzat, T. Poggio, "Visual speech synthesis by morphing visemes," Int J Comput Vis, Vol. 38 pp.45–57, 2000.

[17] S.W. Foo, Y. Lian, "Recognition of visual speech elements using adaptively boosted HMM," IEEE Trans Circuits Syst Video Technol, Vol. 14, pp.693–705, 2004.

[18] A. Rekik, A. Ben-Hamadou, W. Mahdi, "A new visual speech recognition approach for RGB-D cameras," 11th Int. Conf. Image Anal. Recognit., pp. 21–8, 2014.

[19] A. Rekik, A. Ben-Hamadou, W. Mahdi, "Human machine interaction via visual speech spotting," Adv Concepts Intell Vis Syst, pp.566–74, 2015.

[20] V.S. Sadeghi, K. Yaghmaie, "Vowel Recognition using Neural Networks," Int J Comput Sci Netw Secur, Vol. 6, pp.154–8, 2006.

[21] B. Denby, T. Schultz, K. Honda, T. Hueber, J. Gilbert, "Silent Speech Interfaces," Speech Commun, Vol. 52, pp.270–87, 2010.

[22] M. Wand, J. Koutnk, J. Schmidhuber, "Lipreading with long short-term memory," Int. Conf. Acustics, Speech Signal Process., pp. 6115–9, 2016.

[23] T. Hueber, G. Aversano, G. Chollet, B. Denb, G. Dreyfus, Y. Oussar, et al., "Eigentongue Feature Extraction for an Ultrasound-based Silent Speech Interface.," Int. Conf. Acustics, Speech Signal Process., pp. I – 1245 – I – 1248, 2007.

[24] B. Denby, M. Stone, "Speech Synthesis from Real Time Ultrasound Images of the Tongue," Int. Conf. Acustics, Speech Signal Process.," pp. I – 685 – I – 688, 2004.

[25] N. Sugie, K. Tsunoda, "A Speech Prosthesis Employing a Speech Synthesizer – Vowel Discrimination from Perioral Muscle Activities and Vowel Production," IEEE Trans Biomed Eng, Vol. 32, pp.485–90, 1985.

[26] M.S. Morse, S.H. Day, B. Trull, H. Morse, "Use of Myoelectric Signals to Recognize Speech," 11th Annu. Conf. IEEE Eng. Med. Biol. Soc., pp. 1793–17, 1989.

[27] T. Schultz, M. Wand, "Modeling Coarticulation in Large Vocabulary EMG-based Speech Recognition," Speech Commun, Vol. 52, pp.341–53, 2010.

[28] M. Wand, M. Janke, T. Schultz, "Tackling Speaking Mode Varieties in EMG-based Speech Recognition," IEEE Trans Biomed Eng, Vol. 61, pp.2515–26, 2014.

[29] M.J. Fagan, E. SR, J.M. Gilbert, E. Sarrazin, P.M. Chapman, "Development of a (Silent) Speech Recognition System for Patients Following Laryngectomy," Med Eng Phys, Vol. 30, pp.419–25, 2008.

[30] A. Bastanfard, M. Aghaahmadi, A. Kelishami, M. Fazel, M. Moghadam, "Persian viseme classification for developing visual speech training application advances in multimedia information processing," In: Muneesawang P, Wu F, Kumazawa I, Roeksabutr A, Liao M, Tang X, editors. Lect. notes Comput. Sci, Berlin: Springer, , Vol. 5879, pp.1080–5. 2009.

[31] A. Bastanfard, M. Fazel, A.A. Kelishami, M. Aghaahmadi, "A comprehensive audio-visual corpus for teaching sound persian phoneme articulation," 2009 IEEE Int. Conf. Syst. Man Cybern., San Antonio, TX, USA: 2009.

[32] A. Bastanfard, M. Fazel,A. Kelishami, M. Aghaahmadi, "The Persian linguistic based audio-visual data corpus, AVA II, considering coarticulation," Advances in multimedia modeling, In: S. Boll, Q. Tian, L. Zhang, Z. Zhang, Y.P. Chen, editors, Lect. notes Comput. Sci., Berlin: Springer, Vol. 5916, 2010.

[33] K. Noda, Y. Yamaguchi, K. Nakadai, H.G. Okuno, T. Ogata, "Lipreading using convolutional neural network," 15th Annu. Conf. Int. Speech Commun. Assoc., pp. 1149–53, 2014.

[34] Y. Pei, T.K. Kim, H. Zha, "Unsupervised random forest manifold alignment for lipreading," IEEE Int. Conf. Comput. Vis., pp.129–36, 2013.

[35] H. Sakoe, S. Chiba, "Dynamic programming algorithm optimization for spoken word recognition," IEEE Trans Acoust Speech Signal Process, Vol. 26, pp.43–49, 1978.

[36] A.A. Shaikh, D.K. Kumar, W.C. Yau, M.C. Azemin, J. Gubbi, "Lip reading using optical flow and support vector machines," IEEE 3rd Int. Congr. Image Signal Process, pp.327–30, 2010.

[37] G. Zhao, M. Barnard, M. Pietikainen, "Lipreading with local spatiotemporal descriptors," IEEE Trans Multimed, Vol. 11, pp.1254–65, 2009.

[38] D. Yu, "The Application of Manifold based Visual Speech Units for Visual Speech Recognition," PhD Thesis, Dublin City University, 2008.

[39] K. N. Gurney, "Neural Networks," Taylor & Francis Group, 1997.

[40] A.V. Holden, V.I. Kryokuv, "Neural networks - Theory and architecture," NewYourk: manchester university press, 1990.

[41] M.B. Menhaj, "Fundumentals of Neural Networks," Tehran: Amir Kabir University, 2002.

[42] O. Omidvar, J. Dayhoff, "neural network and pattern recognition," London: Academic Press, 1998.

[43] P.J. Werbos, "Beyond regression: New tools for prediction and analysis in the behavioral sciences," PhD Thesis, Cambridg, MA:Harvard, 1974.

[44] T. Shinchi, Y. Maeda, K. Sugahara, R. Konishi, "Vowel recognition according to lip shapes using neural networks," IEEE World Congr. Comput. Intell., 1998.