JACET

J. ADV COMP ENG TECHNOL, 5(2) Spring 2019

Designing and Implementing a System for Automatic Recognition of Persian letters by Lip-reading Using Image Processing Methods

Masoud Barkhan[1],Fattah Alizadeh[2],Vafa Maihami[3]

1-Computer Dept, Technical and Engineering Faculty, Islamic Azad University Sanandaj Branch, Sanandaj, Iran (m_barkhan@yahoo.com)

2-Computer Dept, Technical and Engineering Faculty, Islamic Azad University Mahabad Branch, Mahabad, Iran

3-Computer Dept, Technical and Engineering Faculty, Islamic Azad Unversity Sanandaj Branch, Sanandaj, Iran

Received (2018-11-18) Accepted (2019-02-16)

Abstract: *For many years, speech has been the most natural and efficient means of information exchange for human beings. With the advancement of technology and the prevalence of computer usage, the design and production of speech recognition systems have been considered by researchers. Among this, lip-reading techniques encountered with many challenges for speech recognition, that one of the challenges being the noise in some situations, which is the main cause of errors in the correct diagnosis of speech. One of the ways for solving this problem is image processing, that in this study, the purpose has been designing and implementing a system for automatic recognition of Persian letters through image-processing techniques. For this purpose, after providing a database for Persian verbal phonetics, we first used image processing techniques to eliminate the presence of noises and detect the cantor in lip, in which we used edge detection to identify the edges of the lip. After finding the upper and lower points of the lip for five frames of each film, we used the mean gap between the upper and lower points of the lip as the characteristic of each phoneme and then by providing a database of these features, with the help of the back propagation artificial neural network and The radial basis function have categorized these phonemes, which ultimately achieved the desired results in the categorization of the phonemes. Of course, the precision of classification using the back propagation artificial neural network has been more than radial basis function artificial neural network.*

Keywords: *Automatic letter detecting system, Persian language, lip reading, Image processing.*

How to cite this article:
Masoud Barkhan,Fattah Alizadeh,Vafa Maihami. Designing and Implementing a System for Automatic Recognition of Persian letters by Lip-reading Using Image Processing Methods. J. ADV COMP ENG TECHNOL, 5(2) Spring : 71-80

1. INTRODUCTION

Speech is the most natural and most efficient means of information exchange for human being. In the last half century, designing and production of recognition systems have been the target of research centres. From a long time ago, lip reading was a technique for speech detection and communication. Lip reading techniques have many challenges. In recent years, researchers have been putting these challenges into the computer world and trying to make detect and exchange-spoken information be easier and more accurate by machines as much as possible. In parallel with these advances, recognition of acoustic patterns and signal processing by the machine has a significant help in identifying words and processing information. So far, there have been many advances in voice recognition, however, many devices and software have been implemented; but some of the limitations of these methods have led researchers to look for other ways to complete this domain of data exchange. However, it has been developed

a widespread domain for the Human and Computer Interaction (HCI) that called Visual Speech Recognition (VSR).

The combination of visual methods with audio methods is called audio-video speech recognition (AVSR). Audio techniques are relying on the reception of audio signals that have separated in air, otherwise visual methods (VSR) involve image processing, artificial intelligence, object detection, pattern recognition and so on. A VSR system includes image reception, lip identification, feature extraction, and speech recognition. The precision of a VSR system is highly dependent on the identification of the lips and its limits. In this research, we try to use one of the optimized available methods to identify face and lips and then extract the features and recognize the voice for the Persian language. We know that the voices, how they are pronounced, number and their shapes in various languages are different. Image recognition of voices by the machine for various languages are expanding but this has not been done in many languages, including Persian, or it is seldom. In this paper we attempt to extract the disadvantages and advantages of these methods after applying existing methods and, using image processing techniques, presents a new method for detecting these voices and in a software system (which will be designed and implemented for this purpose) evaluated and tested.

2. BACKGROUND

Jadczyk & Ziolko in a research entitled "The Polish-language audio-video speech processing system, with Dynamic Bayesian network models," describes a speech processing system for Polish language, which uses both audio and video features, and is based on dynamic Bayesian network models. The imaginary nature of information will be extract from the lips of speaker that alternatively is based on discrete cosine transform or features of active appearance model. The acoustic nature was extracted using two standard MFCC parameters and discrete wavelet transform characteristics (DWT). The designed frequency sub-bands were upgraded in accordance with human understanding. An asynchrony between the acoustic and visual nature of the dynamic Bayesian model is shown, too. The system is based on a collection of Polish-language audio and video speeches, which includes the most commonly used phrases in Polish-language conversations, that expressed by 24 speakers. Tests under different busy and quiet conditions confirm that the use of acoustic and visual nature of raw pixel and discrete cosine transform, or the features of the active appearance model of the dynamic Bayesian network, can decrease the rate of error detection of phrases in comparison with standard audio results, the hidden Markov model with MFCC parameterization to 35% in signal-to-noise conditions [1].

Sreekanth & Narayanan, introduced a new method to improve the accuracy of automatic speech recognition system by adding non-vocal parameters. Hint features, which commonly present with speaking are used to improve ASR system accuracy in crowded environments. Both dynamic and static hints are integrated with Speech Recognition System and have been tested in noisy locations. The accuracy of speech recognition system is steady and the system of detection distinct words with gestures and without gestures are tested under various crowded conditions. Their results showed that the addition of gestures features provides sustainability detection accuracy under crowded environments for audio signals [2].

Thanda & Venkatesan presented a learning algorithm for the audio-visual speech recognition automatic system (AV-ASR) using the Deep Reversal Neural Network (RNN). First, they trained a deep RNN vocal model with a connectionist temporal classification objective function (CTC). The framework labels obtained from this audio model were later used to implement the nonlinear dimensional diminishing of the gesture features by deep bottleneck network and the integrated audio and video features to train a combined RNN network. Using the bottleneck features for modality helps that this model properly has had convergence throughout the training. Their results showed that the presence of visual aspect, resulted in significant improvement in the character error rate (CER) at different levels of noise, even when this model trained with any noisy data. They also compared these two methods of integration (integration of features and decision-making integration) [3].

Garg & verma in an investigation entitled "An improved Visualized recognition of English Language Phonemes Using lip-reading technique" introduced two phenomena in emergence of this new field of research. Main section of this discussion is mapping acoustic reformation to visual and the challenge will be accurate and sufficient motion to transmit useful information to the listener in a real-time system with low latency [4].

Lalitha & Thyagharajan in a study entitled "Examining the techniques for positioning the lip for lip-reading in a video" have discussed some of the different ways to determine the position of the lips from the face. Determining the lip position is the main needed step for lip-reading to extract visual information from video. These techniques can be applied to the asymmetric lips, as well as to the visible mouth and teeth, and tongue and mouth with mustache. Generally, in the process of lip-reading, first the location of the lip is determined in the first frame of the video, then the lips are followed by using the pixel points obtained at the initial stage in the subsequent frameworks, and in the end turning the followed lip model into their corresponding matching words to obtain visual information. A proposed new model with the discussed technique is introduced, too. When voice is not available or it is low or accompanied by noise, lip-reading in communication systems is useful for automatic speech recognition. Human-computer communication also requires speech recognition [5].

3. SUGGESTED METHOD

lip-reading based on Image processing methods is one of the interesting topics to researchers, and so far has been done in some languages and also standard databases have been provided, but unfortunately, a standardized database has not introduced for Persian language. Therefore, according to the subject of this research, at first it is necessary to prepare a database. In selecting voices of this database, we should note that some of the voices are articulated in mouth and some in the pre-mouth and throat, the easiest voices in terms of reading, are labial voices (on the lip) and dental voices [6]. Therefore, according to this point, the selected labial voices are according to Table 1.

Table 1. The selected labial voices for the database

آ	د
اُ	ف
اِ	ل
او	م
ای	ن
ت	و

In order to taking picture, a Samsung camera (480 × 640 VGA) has been used. The reason for choosing this image resolution is that the suggested method can work on low-quality images that are received from conventional CCTV cameras, whilst image processing with less resolution takes less time. Finally, the collected database includes labial voices that articulated by 11 people (8 women and 3 men), which was screened in terms of quality and noise, and finally 132 videos have been selected. The number of frames captured from the voices is not the identical. Therefore, it's not possible to use all frames for images, so we will only extract a number of frames. With the reviews that we've done, the middle frame of the videos, clearly shows the photo of each voice. So we used the following series to select several frames. The reason for not choosing successive frames is that the changes in the images are inconsiderable.

$$T = \{n\text{-}4, n\text{-}2, n, n\text{+}2, n\text{+}4\} \quad (1)$$

Where n is equal to, 2 times the number of frames. To make a fully automated system that can determine the location of the lips, first requires locating the face. We use the Viola Jones algorithm to recognize face. One of the most used and powerful face detection algorithms designed by Paul Viola and Michael Jones. To identify objects, they introduced a machine-based learning approach, in which by combination of a large number of poor learners, a strong classifier system is taught. For a problem, that requires only two categories and with labelled exercises samples, an Adaboost learning algorithm (Adaptive Boosting or Adaboost) can select a small number of visual attributes to provide the most accurate classification [7]. In order to

prevent and reduce the effects of unwanted light changes from environmental factors, we use the equation (2) to compensate for the changes [7].

$$g(x,y) = k + \frac{log(f(x,y)+1)}{d * log(t)} \qquad (2)$$

Where *f(x, y)* is the original image and *g(x, y)* is the image obtained from the compensation for the Luminance changes. By performing the test on various images, the values obtained for the coefficients t, d and k are 2, 0.25, and 10 respectively. The colored images of the database that we used in our study for analysis are in RGB format. Because the distance between the two points in the color space is a factor for understanding the difference among them, so we need a uniform color space to use this space in direction of our goals. The color space CIE L*a*b* has a uniform feature. Perceptual uniformity shows how the two colors for a human are different in appearance. The color space of the Luminance component L^1 and the Chroma component $(ab)^2$ are obtained by a non-linear mapping of the XYZ coordinates. Another important reason for choosing this space and using the feature vectors {L*, a*, b*} is to increase the resistance of the segmentation for a wide range of lip and skin colors. With the respect of mentioned reasons this space was used. In figure 1, the original image and the obtained feature vectors from this image are shown.

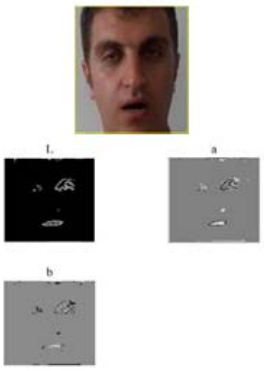

Fig. 1. Original image and color space components CIE L*a*b* derived from this image

As you can see, the resolution of lips area in feature vector is greater than other feature vectors. Then we use feature vector to follow the other steps of image processing.

Fig. 2. Sample of derived lips area

Edge detection has been used to find corners of the lip. In fact, the suggested method focuses on this area. In practice, because of the noise that breaks down the uniform projection of the edges, and other factors that create false fractures in intensity, the resulting pixels are less discriminate of the edge. Usually, after detection algorithms, connecting methods are used to incorporate meaningful pixel edges. One method that can be used to find and connect line segments in an image is the Hough transform. In Figure 2, the corners of the lip found with the edge detection method for the phoneme "ï" are shown.

After the steps of preprocessing the images, it's time for the contour of the lips. To find the contour of the lips, we use the proposed algorithm by [8]. Using the line detection with Hough transformation, we could find the outstanding points of the outer contour edges of the lip and, by calculating the distance between the two lines, we find the midpoint of the lip contour. Using the midpoint of the lip contour, we obtained the points of the upper and lower midpoint of the lip contour. According to equations (3) to (6), we obtained the remaining features of the contour of the lips; the result is in the form of Figure 4.

$$x_c = {}^1\!/_2\ (la_x + lb_x)\ ,\ y_c = {}^1\!/_2\ (la_y + lb_y) \qquad (3)$$

$$a = {}^1\!/_2\ ((la_x - lb_x)^2 + (la_y + lb_y)^2)^{1/2} \qquad (4)$$

$$b_{up} = ((va_x - x_c)^2 + (va_y - y_c)^2)^{1/2} \qquad (5)$$

$$b_{low} = ((vb_x - x_c)^2 + (vb_y - y_c)^2)^{1/2} \quad (6)$$

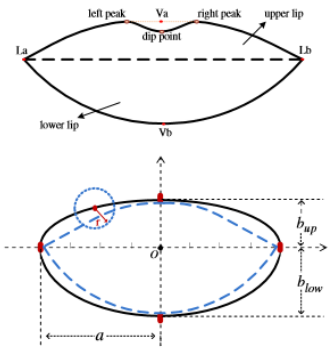

Fig. 4. Derivation of lip contour

In this algorithm, a 16-point model is used to describe the lip contour, and the lip margin points are labeled in the clockwise direction in Fig. 5. The set of parameters of the lip model can be described by $\lambda_p = x_{pi}, y_{pi}$, i= 1,2,3,…,16 . These points are divided into two groups: $P_9 - P_{16}$ and P_1 describe lower lip and $P_9 - P_{16}$ describe upper lip. P_1 and P_9 exhibit corners of lip and P_5, is concavity point. The corners of the lip are obtained by edge detection technique and the other 14 points of the lip using the color intensity. The x coordinates of these points are defined on the x-axis, for example, the x coordinates of the points between the corners of the lip are equal to $\left\{-\frac{3}{4}, -\frac{1}{2}, \frac{1}{4}, 0, \frac{1}{4}, \frac{1}{2}, \frac{3}{4}\right\}$.

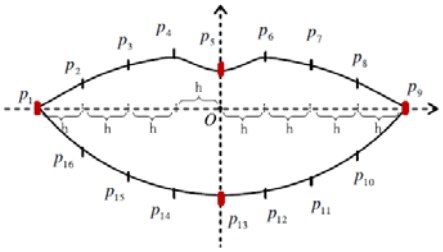

Fig. 5. Sample of derived lip area

In Fig. 6 the sample of derived lip area for "ĭ" phoneme is shown.

Fig. 6. Sample of derived lip area for "ĭ" phoneme.

After finding the contour of the lips, it is feature extraction time. The features based on the distance between the points and the angle of the corners of the lip have considered. Namely the intervals between P_1 to P_9, P_2 to P_{16}, P_3 to P_{15}, P_4 to P_{14}, P_5 to P_{13}, P_6 to P_{12}, P_7 to P_{11}, P_8 to P_{10}, and the angle between P_2-P_1-P_{16} and P_8-P_9-P_{10} Extract as the features of the lip. Then entirely 10 features of the Contour of Lips are extracted from the images for each phoneme. As regards, each phoneme is 5 frame, and from each frame, 10 features are extracted, so for each phoneme, 50 features have extracted. To reduce these features and to find better results in categorizing the voices, we use an average of 10 frames.

In this study, we used the back propagation artificial neural network and radial base to categorize the voices. The aim is to compare these two artificial neural networks and finding the best method for categorizing Persian language voices. We divide the vector of inputs of artificial neural networks randomly into three sets as follow:

- 60% of the data for the educational set;
- 20% of the data in the evaluation set in order to prevent network over fitting;
- 20% of the data in the test set to check the performance of the final network.

The most prevalent architecture of back propagation networks is the feed-forward multi-layered networks. The first step in designing this network is the creation of a network object. The newff function creates a feed forward network. This function has three inputs and returns the supplied network as output. The input parameters of this network are: input vector, target vector and an array containing hidden layers. In this type of artificial neural network, the default parameters are used, which are summarized in Table 2.

Table (2). Parameters of back propagation Artificial Neural Networks

trainlm	Training Function
learnpn	Learning Function
Tansig, logsig , purelin	Transfer Function

For radial basis function neural networks, the *newrb* function is also used. The *newrb* function continuously adds neuron to basis radius network. This will continue as long as the errors of sum square be less than the defined goal value or we have reached the maximum value of the specified neuron value. The *newrb* function receives the values of P(*input matrix*), T(*target matrix*) and design parameters GOAL and SPREAD as inputs and returns the optimal network as output. In this type of artificial neural network, target functions *Radbas* and *Pureline* have been used. In Table 3 and 4, the inputs and outputs of artificial neural networks are shown.

Table (3). Artificial neural network inputs

Input	Description
a	from P1 to P9
b	from P5 to P13
L1	from P2 to P16
L2	from P3 to P15
L3	from P4 to P14
L4	from P6 to P12
L5	from P7 to P11
L6	from P8 to P10
Z1	Angle P2-P1-P16
Z2	Angle P8-P9-P10

Table 4. Artificial neural network outputs

Phoneme	Category	Phoneme	Category
أ	7	آ	1
إ	8	او	2
ت	9	ای	3
د	10	ف	4
م	11	ل	5
ن	12	و	6

4. RESULTS

According to the research, there are no standards for determining the number of neurons and hidden layers in artificial back propagation neural networks, rather they obtained randomly. Therefore, in this paper, we use the trial and error method to determine the best number of neurons and hidden layers. The test results are shown in Tables 5 and 6.

Table 5. The performance of first proposed back propagation artificial neural network based on the number of hidden neurons

number of hidden neurons	back propagation artificial neural network error
1	0.49974
2	0.38900
3	0.38353
4	0.29816
5	0.52201
6	0.54244
7	0.68648
8	0.55654
9	0.61574
10	0.48854

According to Table 5, the least error is related to 4 neurons in the hidden layer, so the best back propagation artificial neural network to classify the voices contains a hidden layer with the number of 4 hidden neurons.

Table 6. The performance of first proposed radial basis function Artificial Neural Network Performance Based on the number of hidden neurons

number of hidden neurons	radial basis function Artificial Neural Network error
1	0.406203
2	0.378312
3	0.337164
4	0.357498
5	0.365678
6	0.324625
7	0.333697
8	0.359116
9	0.326391
10	0.325194

According to Table 6, the least error is related to 6 neurons in the hidden layer, so the best radial basis function artificial neural network to classify the voices contains a hidden layer with the number of 4 hidden neurons.

Fig. 7 shows the performance diagram of the back propagation artificial neural network during the training process. The training routine will be stopped if the evaluation set error in 6 consecutive repetition raised; that this stoppage has occurred in the eleventh repetition. The best validation for this group of data in the fifth cycle is 0.29816.

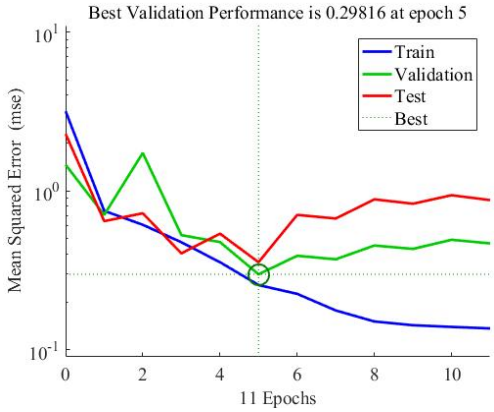

Fig. 7. The performance of suggested back propagation neural network

Fig. 8 shows the performance diagram of the radial basis function artificial neural network during the training process. As can be seen, during the training process and during 6 repetitions, the training error has decreased, which in the end, in the sixth repetition, the performance of the radial basis function artificial neural network has reached to 0.324625.

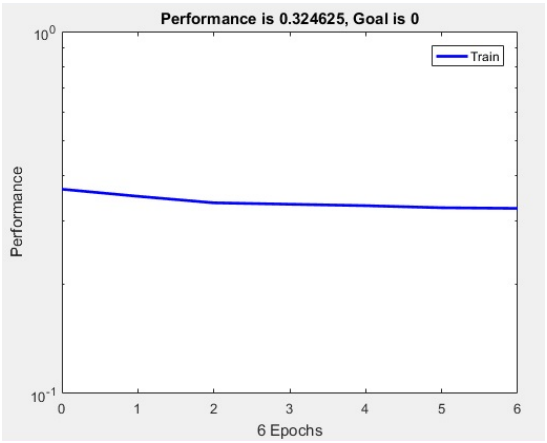

Fig. 8. The performance of suggested radial basis function artificial neural network

Generally, with the training of back propagation artificial neural networks and the radial basis function, we can reduce the error and categorize the voices. However, by comparing the performance of back propagation artificial neural networks and the radial basis function, it can be concluded that the back propagation neural network performance is better than the radial basis function. One of the criteria for evaluating the performance of artificial neural networks is the calculation of the mean square error for training, testing and evaluation data. The results are shown in Table 7.

Table 7. Mean square error of proposed artificial neural networks

Back propagation	training	0.2466
	trial	0.4711
	assessment	0.5732
Radial basis function	training	0.3246
	trial	0.4937
	assessment	0.6636

The error of the training, testing, and evaluation data of the artificial neural network is lower than radial basis function, so it can be concluded that back propagation artificial neural network with less error has been able to

better classify voices. For the accuracy of these results, Tables 8 and 9 describe the results of the classification of the voices.

Table 8. The Results of classification of voices by the proposed radial basis function artificial neural network

Phoneme	correct	incorrect	percent	Phoneme	correct	incorrect	percent
آ	10	1	91	أ	7	4	63.5
او	9	2	81.82	إ	9	2	81.82
ای	7	4	63.64	ت	9	2	81.82
ف	9	2	81.82	د	9	2	81.82
ل	9	2	81.82	م	8	3	72.73
و	7	4	63.64	ن	6	5	54.56

Table (9). The Results of classification of voices by the proposed back propagation artificial neural network

Phoneme	correct	incorrect	percent	Phoneme	correct	incorrect	percent
آ	10	1	91	أ	7	4	63.64
او	10	1	91	إ	10	1	91
ای	8	3	72.73	ت	10	1	91
ف	10	1	91	د	10	1	91
ل	10	1	91	م	9	2	81.82
و	9	2	81.82	ن	7	4	63.64

So, in general, it can be concluded that the accuracy of the classification of the voices by the back propagation artificial neural network is higher than the radial basis function. For better and more investigation about the proposed method, we started to collect a new database. In this new database, we set the imaging quality to HD 720P, and made 10 videos (2 women and 8 men) of the labial voices. After applying the pre-processing of images that described in chapter 3 and extracting the desired features. In Fig. 9, an example of the extraction of the contour of the lip is shown on the new database images. As you can see, with increased image quality, the number of pixels in the images increased, and the extraction of the lip contour was made more accurately.

Figure (9). Extracted Lip Contour for an image with HD 720p quality

Mean square error for training, testing and evaluation data are shown in Table 10.

Table (10). Mean square error second proposed artificial neural networks

Back propagation	training	1.3331e-14
	trial	3.3735e-7
	assessment	9.3357e-9
Radial basis function	training	0.0069
	trial	0.0063
	assessment	0.0554

The error of the training, testing and evaluation data of the back propagation function artificial neural network is lower than the radial basis function, so it can be concluded that the back propagation artificial neural network of with less error has better performance to classify the voices. For the accuracy of these results, we describe the results of the classification of the voice in Table 11 and 12.

Table 11. The classifying results of the second proposed radial basis function artificial neural network

Phoneme	correct	incorrect	percent	Phoneme	correct	incorrect	percent
آ	9	1	90	أ	8	2	80
او	9	1	90	إ	9	1	90
ای	6	4	60	ت	9	1	90
ف	9	1	90	د	8	2	90
ل	9	1	90	م	8	2	80
و	8	2	80	ن	5	5	50

Table (12). The classifying results of the second proposed back propagation artificial neural network

Phoneme	correct	incorrect	percent	Phoneme	correct	incorrect	percent
آ	10	0	100	أ	9	1	90
او	10	0	100	إ	10	0	100
ای	9	1	90	ت	10	0	100
ف	10	0	100	د	9	1	90
ل	10	0	100	م	9	1	90
و	9	1	90	ن	8	2	80

5. CONCLUSION

In this paper, we tried to introduce a methodology for designing an automatic system for lip-reading so that the system can be used to recognize articulated vowels and consonants. In this regard, we conducted a comprehensive study on the existing methods in this field and introduced a proposed method. For this purpose, we first gathered the database. Finally, we introduced two databases of 480 × 640 VGA and HD 720P. Then, by using of image processing, we did preprocessing on images extracted from video files. Finally, we extracted the contour of the lip and extracted 10 features from each image. The extracted features were exerted as inputs to the back propagation artificial neural network and the radial basis function to categorize the voices. In the end, we achieved to the following results:

- As the quality of image increases, the image pixels will increase and the extraction of the lip cantor will be more accurately.
- Extracting HD 720P database image features are more accurate than the 480 × 640 VGA database images.
- Back propagation artificial neural network error is less than the radial basis function, so the performance precision of the back propagation artificial neural network is higher than the radial basis function.
- Classification of voices based on back propagation artificial neural network is more accurate than the radial basis function.
- The classifying results of the HD 720P database images are better than the 480 × 640 VGA database images.

The major achievements of this article include:

1- Introducing a new method for better segmentation of the lip area

2- Introducing a new feature to categorize and provide an algorithm to improve system recognition accuracy

3- Using artificial neural networks, which has faster training and responding than previous methods (in prior studies) in this field.

In the future studies, this system can be used to detect vowels and consonants in various speakers. For doing this, you should use the marking on the images of the lips of the different speakers and training the exerted network. In order to increase the accuracy of this system, information of speaker's sound can be used simultaneously along with visual information. For doing this, in addition to the features extracted from the images, a series of other features should be extracted from the voice of the speaker and used in the training of the network; in this case, the number of extracted features may increase, which for better selecting of features, methods such as SFS can be used.

For the reason that other researchers have used a different training database proportionate to their study, we can't make a proper comparison, but each step can be compared separately with other methods of research.

REFERENCES

1. Jadczyk, Tomasz. Zi´ołko, Mariusz. (2015). "Audio-Visual Speech Processing System for Polish with Dynamic Bayesian Network Models". Proceedings of the World Congress on Electrical Engineering and Computer Systems and Science (EECSS).

2. Sreekanth, N. S., Narayanan, N. K. (2016). "Enhanced Automatic Speech Recognition with Non-acoustic Parameters". Proceedings of the International Conference on Signal, Networks, Computing, and Systems pp 93-104.

3. Thanda, Abhinav. Venkatesan, Shankar. (2016). "Audio Visual Speech Recognition using Deep Recurrent Neural Networks". Computer Vision and Pattern Recognition (cs.CV).

4. Garg, Ishu. Verma, Amandeep. (2016). "An improved visual Recognition of letters of English Language Using Lip Reading Technique". Garg Ishu, Verma Amandeep, International Journal of Advance research, Ideas and Innovations in Technology.

5. Lalitha. S.D., Thyagharajan K.K. (2016). "A Study on Lip Localization Techniques used for Lip reading from a Video". International Journal of Applied Engineering Research ISSN 0973- 4562 Volume 11, Number 1, pp 611-615.

6. Khanlari, Parviz. (2015) Dastur Zaban Farsi (Grammar of Persian language)

7. Wang, Yi-Qing., (2014). V0.5 IPOL article class An Analysis of the Viola-Jones Face Detection Algorithm. pp. 128-148.

8. Cheunga, Y., Liua, X., Youb, X., (2012). "A local region based approach to lip trackin", Contents lists available at SciVerse ScienceDirect.