Continuous Persian Phones Recognition Using Lip Reading

Yousef Taheri, Mohammad Hosein Yektaie, Amir Masoud Rahmani
Department of Computer, Science and Research Branch, Islamic Azad University (IAU), Khouzestan, Iran
youtaheri@yahoo.com, mh_yektaie@yahoo.com, rahmani74@yahoo.com

***Abstract*— The main preference in this paper by comparison with other researches is given to application of several specifications together in effect to gait toward creating a suitable system for dumb and deaf usage. In other words, in this paper, we try to recognize continuous phones in different words, uttered by different males and females, with different uttering speed, without using hand-labeled model on lips and without wasting time on processing unnecessary pixels. In addition, we use our best effort to have the most usage of pixel colours. Although most of these specifications cause confusions in recognition, we believe that considering these specifications together in researches, results in creating a real time automatic lip reading system in less time in future. The result of testing 12 different utterances were 66.1% for phones period detection and 51.3% for phones recognition.**

***Keywords*— Lip reading, Continuous phones, K-NN, Persian, Contour.**

I. Introduction

One of the God's gifts that people use to communicate with others is lip. But unfortunately some people are unable to use their lips efficiently. Usually these people can not send their sound to others and can not hear other's speech. Thus, the motion of mouth regions becomes important for their communications, but because healthy people are unable to read their lips, they lay stress on hand gestures and take no notice of the use of lips. If there is a system that enables them to read lips, they'll try to correct their words utterance. By the development of sound processing, there is no problem for deaf people to get others' speech, but in noisy environments or for dumb people to communicate with others, creating lip reading system can be useful.

Wide and good researches have been done in the field of visual and audio-visual lip reading (e.g. [1]), but most of these approaches are not suitable for dumb and deaf people, because some of them carry out one or more steps of the lip reading process by hand (e.g. [2]–[4]), some recognize separated phones or words (e.g. [4]–[6]), some use speaker dependant methods [7], and some others use just one or a few subjects to create or learn their lip reading systems (e.g. [8], [9]). Certainly interesting approaches are used in these researches and they can be used for other purposes such as speaker identification by lip reading, as in [3], or lip reading for counter-terrorism and law enforcement areas (e.g. [10]), etc., but in this paper, was tried to delete hand-performed steps and automate them. Also, we restrained that pixel processing that has a little effect on accuracy.

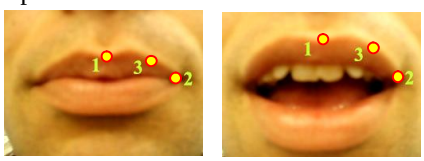
Because of almost high similarity between phones in all languages and knowing that words are constructed by phones, we focused on recognizing phones. Due to the similarity of some phones utterances created by lips and teeth (not tongue), we classified all Persian phones in 16 sets, and we tried to recognize continuous phones in different words uttered with different speeds and by different speakers, including males and females .

II. Locating Major Points

Although using a model for lips contour extraction (e.g. [4]), results in a good care, we have to prepare a hand-labelled model (for more accuracy), which can not be suitable for real time or automatic lip reading.

Regarding the fact that we have attempted to take steps to achieve automatic and real time lip reading, we have not used such hand labelled models.

Moreover, in this paper we locate points of lips that really help us extract useful features of lip reading, and we prevent wasting time on locating useless points on lips. For example, in Fig. 1 that shows the utterance starting and finishing frames of the phone /bæ/ we perceive that when points 1 and 2 moves farther from the gravity of the mouth, point 3 also moves farther. It's impossible that something else happens in somebody's lips.



(a) (b)

Fig. 1. Utterance starting and finishing frames of phone /bæ/

Therefore, in this paper we just locate 5 points on lips as shown in Fig. 2.

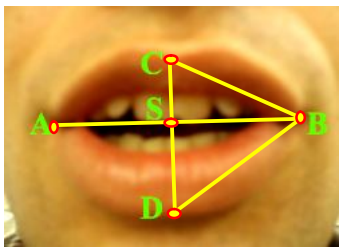


Fig. 2. Locating 5 points on lips

To locate these points, steps of Fig. 3 are done.

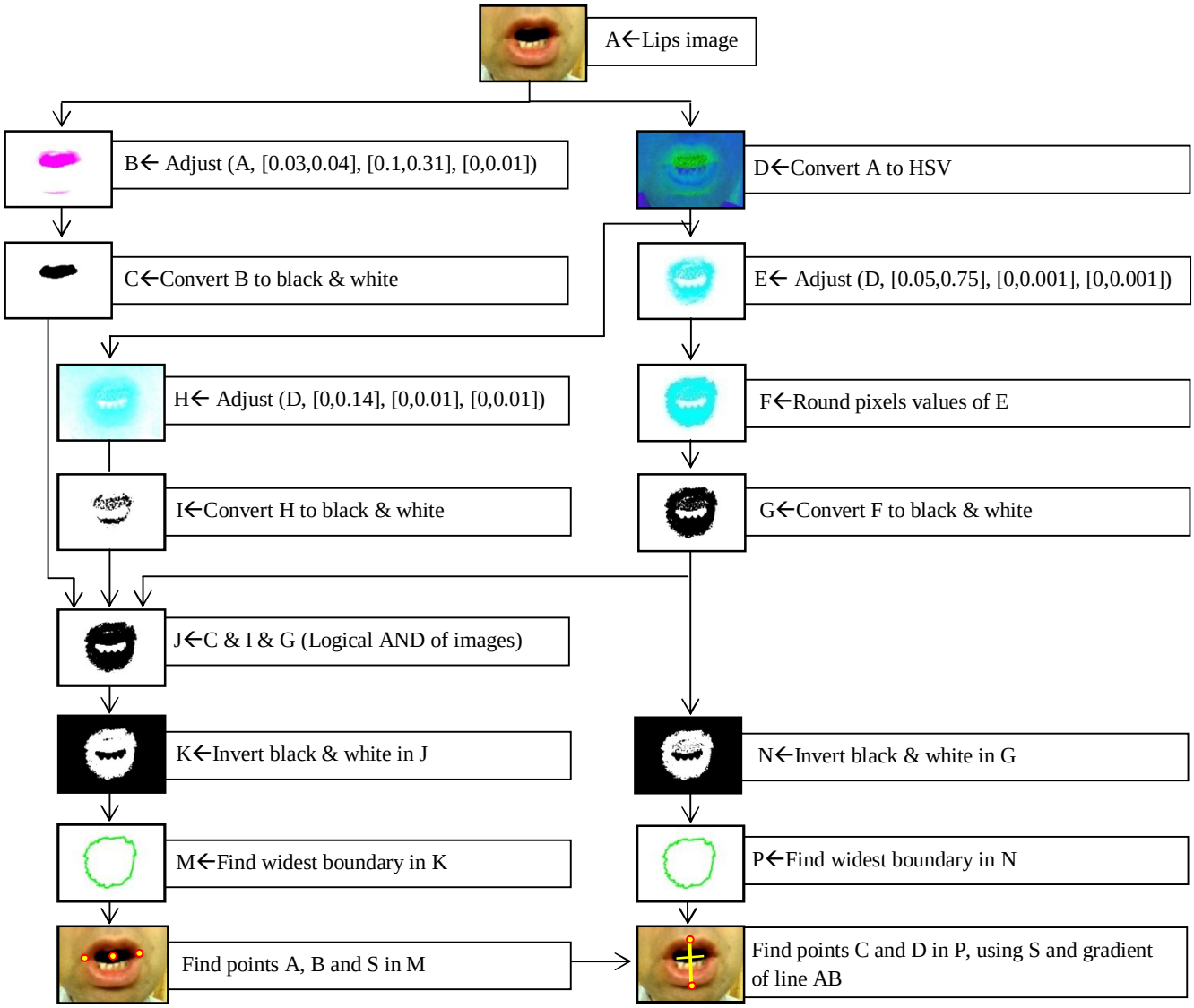


Fig. 3. Locating points A, B, S, C and D on lips

In Fig. 3, the instruction "Adjust(lips,[rl,rh],[gl,gh],[bl,bh])" maps the red values between "rl" and "rh" to values between 0 and 1, and also exerts it on green and blue colours of image.

Values below "rl" and above "rh" are clipped, that is, values below "rl" map to 0, and those above "rh" map to 1. Also we can use this instruction for HSV images.

But locating point D in this method for some cases, on some lips and in some light conditions doesn't work carefully. For example, Fig. 4 shows that point D is located wrongly.

Fig. 4. Wrong location of point D

Therefore, we locate point D using another method that is shown in Fig. 5.

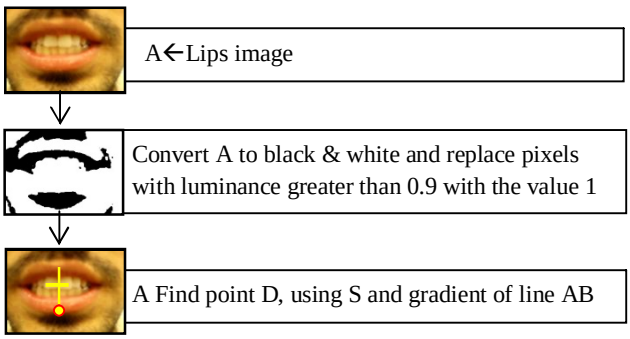


Fig. 5. Another method for locating point D

Although in Fig. 3 some steps look unnecessary for this example, for some other lips in different light conditions we have to apply them.

Moreover, in some cases that the shadow of the lower lip is not created (as shown in Fig. 6), the second method can not be useful, so for each frame we locate point D by both methods and each of them which is nearer to point D of the fore frame is selected as the right location of point D in the present frame.

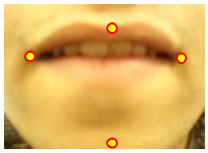


Fig. 6. Wrong location of point D, by using second method

In this paper, in order to use the colour of mouth regions efficiently, we applied RGB and HSV colours several times, and finally we merged them to locate the points.

III. Feature Extraction

A. *Proposed Features*

To take the most advantage of points obtained in II, we defined 8 following features:

1) The whiteness of region CSB to the hole of region CSB ratio. (for teeth appearance)

2) The whiteness of region DSB to the hole of region DSB ratio.

3) The cavity appeared in region CSB to the hole of region CSB ratio.

4) The cavity appeared in region DSB to the hole of region DSB ratio.

5) The length of SB to the length of SC ratio.

6) The length of SB to the length of SD ratio.

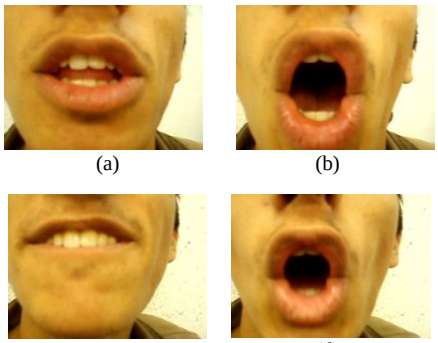
7) The length of SB to the length of SC+SD ratio.

8) The length of SC to the length of SD ratio.

B. *Features Specifications*

The general specifications are given as follows:

- These features do not spend much time to be computed because we used symmetry specification of human lips and we processed just half of the lips for preventing a waste of time. Therefore, in this paper pixel process is at most half of other pixel-based lip readings (e.g. [7]), because we process just pixels of CSB and DSB regions.
- At least 2 features change by the least movement of lips regions.
- Changing the distance of lips from video camera through utterance does not affect the features and does not need a normalization processing for this aim (e.g. [4]), because all of the features are proportional and they are calculated relative to lips size in the same frame.
- The reason for laying stress on separating features of upper and lower lips in this paper, is to distinguish the phones better. For example, Fig. 7 shows the utterance starting and finishing frames of phones /nɑ:/ and /vɑ:/. As observed in Fig. 7(a) and Fig. 7(c), the amount of visible whiteness in both of images are almost the same, but separating features 1 and 2 causes better detection of these phones, because this amount of whiteness in Fig. 7(c) is computed as feature 1, but in Fig. 7(a) this amount is divided between features 1 and 2.



(a) (b) (c) (d)

Fig. 7. Utterance starting and finishing frames of phones /nɑ:/ and /vɑ:/

As another example, see Fig. 8 that shows the utterance starting frame of phones /fe/ and /be/.

(a) (b)

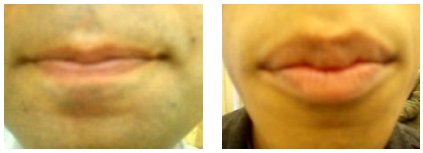
Fig. 8. Utterance starting frame of phones /fe/ and /be/

If we had defined just ratio of height to width of the lips, this ratio in Fig. 8(a) and Fig. 8(b) would have shown almost the same value but by adding features like 5, 6 and 8, these phones are distinguished better. Although in these images the ratio of height to width of lips are almost the same, in Fig. 8(b) the distances of the upper and lower lips from the gravity of lips are almost the same while in Fig. 8(a) the distance of the upper lip is greater than the distance of the lower lip.

Maybe in the first look, by defining these features, feature 7 looks unnecessary but in 4 its advantage will be more cleared up.

C. *Using Criterion Frame*

Fig. 9 shows the utterance starting frame of phone /bɑ:/ by different persons.



(a) (b)

Fig. 9. Utterance starting frame of phone /bɑ:/ by different persons

The value of feature 7 that represents the height to width ratio of lips, is much smaller in Fig. 9(a), than this feature value in Fig. 9(b), while this feature that can be helpful to recognize the same phones uttered by different persons has almost the same value. Therefore, we considered the utterance starting frame of each person that shows the usual and closed state of lips as a criterion frame and replaced the features values by different values and features values of the criterion frame. Hence, the thickness and thinness of lips can not be inconvenient in the phones recognition.

IV. Phones Period Detection

After processing the images to extract the features, we processed the signals drawn by frame features in order to detect the limits of the continuous phones.

A. Normalization

After signal processing we found out that in some points, signals come up and down immediately. By referring to those frames we saw that their image processing was done wrongly because of the wrong location of points B, C, and D. To delete the unsuitable effect of those frames on signal of feature 7, whenever in spite of being descent of the signal, for one frame, feature 7 increased just for a frame, or in spite of being ascent of the signal, for one frame, feature7 decreased, we replaced the features values of such frames by features average values of their pre- and post- frames. Thus, the bad effects of the wrong points located on lips in II are avoided to some extent.

B. Phones Limits Detection

While the target of this paper was the recognition of continuous phones with different uttering speed, we tried to detect the period of the phones automatically by using features signals. After processing the signals and their various combinations, we concluded that feature 7 can be helpful for separating the frames of the phones, because through each phone utterance, the height to width ratio becomes smaller. Hence, in Fig. 10 that shows the signal of feature7 for uttering some Persian words, decreasing the signal shows the utterance starting and finishing frames of the phone.

C. Reducing Fore Phone Effect

Fig. 11 shows that the utterance starting frame of phone /dɑ:/ belongs to one person, but before Fig. 11(a), he uttered phone /fæ/, while before Fig. 11(b), he uttered phone /khəʊ/, and although both phones are one, the lips states are different

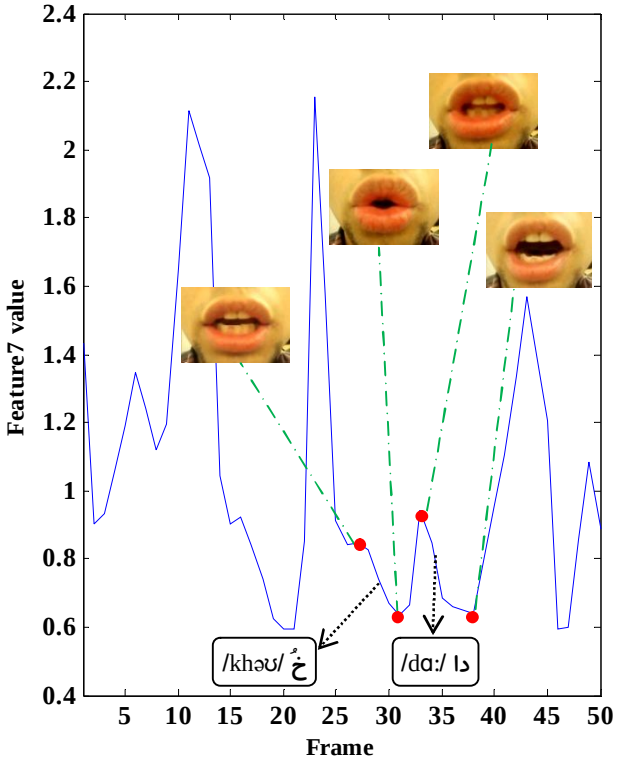
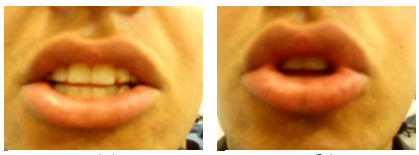


Fig. 10. Signal of feature 7 for uttering some Persian words including 'khoda'

in these images.

To reduce the fore phone effect, we overlooked one-fourth of the starting frames of each phone.



(a) (b)

Figt. 11. Two utterance starting frames of the phone /dɑ:/

V. Recognition

Various methods are used in lip reading in other papers. One of them is Least Mean Square (LMS) (e.g. [11]), but we did not get mentionable recognition rate by using it, because we wanted to train different lips features to be able to work on different lips efficiently and in spite of using criterion frame (as mentioned in III-C), the extracted features of individual phone images have different values for different persons' lips.

Two other methods are hidden Markov model (HMM) (e.g. [3], [4], [8]), and Dynamic Programming (DP) matching (e.g. [5], [7]), but while we intend to recognize phones in this paper, they can not be useful, because unlike the words, for the phones, features values do not increase and decrease through the utterance, and it is just two different states (utterance starting and finishing frame), and middle frames have values between them.

At last, between Neural Network (e.g. [12], [13]), and K Nearest Neighbor (K-NN) we chose K-NN method, because the lips are trained and tested faster. Moreover, for phones recognition as compared with words recognition (e.g. [5], [11]), the importance of feature preparation is much more than recognition phase. Thus, the scrupulous for choosing and comparing the recognition methods can not affect the result very much. Regarding more accuracy, to recognize each phone we used K-NN method twice (as shown in Fig. 12), once for the utterance starting frame and the other time for the utterance finishing frame. We did not exploit other frames because the features values of the middle frames increase and decrease monotonously.

In addition, the feature signal slope of the middle frames can help us in phones recognition process because the utterance speed also causes a change in feature signal slope and regarding the fact that the speed of the words utterance at the time of taking films was different, this difference of signal slope can not be mentioned as a feature or specification of a phone.

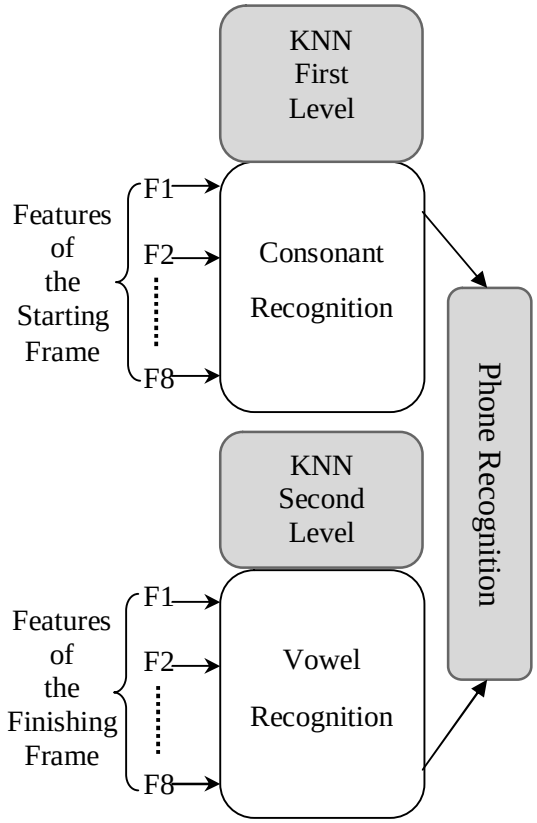


Fig. 12. The applied method of the phone recognition

VI. Experiment

Due to the similarity of some phone utterances in Persian language, we classified all vowel phones into 16 groups. In other words, homophonous specification was considered in this classification. Table 1 shows the phones that are classified in these groups.

TABLE 1
Classification of Phones Into 16 Groups

Group1	Group2		Group3		Group4
/ɑ:/ آ	/u:/ او	/əʊ/ اُ	/i:/ ای	/e/ اِ	/æ/ اَ
/tɑ:/ تا	/tu:/ تو	/təʊ/ تُ	/ti:/ تی	/te/ تِ	/tæ/ تَ
/hɑ:/ حا	/hu:/ حو	/həʊ/ حُ	/hi:/ حی	/he/ حِ	/hæ/ حَ
/khɑ:/ خا	/khu:/ خو	/khəʊ/ خُ	/khi:/ خی	/khe/ خِ	/khæ/ خَ
/dɑ:/ دا	/du:/ دو	/dəʊ/ دُ	/di:/ دی	/de/ دِ	/dæ/ دَ
/rɑ:/ را	/ru:/ رو	/rəʊ/ رُ	/ri:/ ری	/re/ رِ	/ræ/ رَ
/tɑ:/ طا	/tu:/ طو	/təʊ/ طُ	/ti:/ طی	/te/ طِ	/tæ/ طَ
/ɑ:/ عا	/u:/ عو	/əʊ/ عُ	/i:/ عی	/e/ عِ	/æ/ عَ
/ghɑ:/ غا	/ghu:/ غو	/ghəʊ/ غُ	/ghi:/ غی	/ghe/ غِ	/ghæ/ غَ
/qɑ:/ قا	/qu:/ قو	/qəʊ/ قُ	/qi:/ قی	/qe/ قِ	/qæ/ قَ
/kɑ:/ کا	/ku:/ کو	/kəʊ/ کُ	/ki:/ کی	/ke/ کِ	/kæ/ کَ
/gɑ:/ گا	/gu:/ گو	/gəʊ/ گُ	/gi:/ گی	/ge/ گِ	/gæ/ گَ
/lɑ:/ لا	/lu:/ لو	/ləʊ/ لُ	/li:/ لی	/le/ لِ	/læ/ لَ
/nɑ:/ نا	/nu:/ نو	/nəʊ/ نُ	/ni:/ نی	/ne/ نِ	/næ/ نَ
/hɑ:/ ها	/hu:/ هو	/həʊ/ هُ	/hi:/ هی	/he/ هِ	/hæ/ هَ
/jɑ:/ یا	/ju:/ یو	/jəʊ/ یُ	/ji:/ یی	/je/ یِ	/jæ/ یَ
Group5	**Group6**		**Group7**		**Group8**
/θɑ:/ ثا	/θu:/ ثو	/θəʊ/ ثُ	/θi:/ ثی	/θe/ ثِ	/θæ/ ثَ
/dʒɑ:/ جا	/dʒu:/ جو	/dʒəʊ/ جُ	/dʒi:/ جی	/dʒe/ جِ	/dʒæ/ جَ
/tʃɑ:/ چا	/tʃu:/ چو	/tʃəʊ/ چُ	/tʃi:/ چی	/tʃe/ چِ	/tʃæ/ چَ
/ðɑ:/ ذا	/ðu:/ ذو	/ðəʊ/ ذُ	/ði:/ ذی	/ðe/ ذِ	/ðæ/ ذَ
/zɑ:/ زا	/zu:/ زو	/zəʊ/ زُ	/zi:/ زی	/ze/ زِ	/zæ/ زَ
/ʒɑ:/ ژا	/ʒu:/ ژو	/ʒəʊ/ ژُ	/ʒi:/ ژی	/ʒe/ ژِ	/ʒæ/ ژَ
/sɑ:/ سا	/su:/ سو	/səʊ/ سُ	/si:/ سی	/se/ سِ	/sæ/ سَ
/ʃɑ:/ شا	/ʃu:/ شو	/ʃəʊ/ شُ	/ʃi:/ شی	/ʃe/ شِ	/ʃæ/ شَ
/sɑ:/ صا	/su:/ صو	/səʊ/ صُ	/si:/ صی	/se/ صِ	/sæ/ صَ
/ðɑ:/ ضا	/ðu:/ ضو	/ðəʊ/ ضُ	/ði:/ ضی	/ðe/ ضِ	/ðæ/ ضَ
/ðɑ:/ ظا	/ðu:/ ظو	/ðəʊ/ ظُ	/ði:/ ظی	/ðe/ ظِ	/ðæ/ ظَ
Group9	**Group10**		**Group11**		**Group12**
/bɑ:/ با	/bu:/ بو	/bəʊ/ بُ	/bi:/ بی	/be/ بِ	/bæ/ بَ
/pɑ:/ پا	/pu:/ پو	/pəʊ/ پُ	/pi:/ پی	/pe/ پِ	/pæ/ پَ
/mɑ:/ ما	/mu:/ مو	/məʊ/ مُ	/mi:/ می	/me/ مِ	/mæ/ مَ
Group13	**Group14**		**Group15**		**Group16**
/fɑ:/ فا	/fu:/ فو	/fəʊ/ فُ	/fi:/ فی	/fe/ فِ	/fæ/ فَ
/vɑ:/ وا	/vu:/ وو	/vəʊ/ وُ	/vi:/ وی	/ve/ وِ	/væ/ وَ

To obtain training sets, we used 7 word sets that are uttered by different people (males and females). These sets include phones from all 16 groups. Then we extracted different

phones from the continuously uttered words as training sets.

To test our system different males and females uttered different continuous words with different speeds 12 times and videos that were taken by a digital video camera were converted to 250×180 pixels images.

After the feature computations for all images, we detected period of phones (as explained in IV). Then as mentioned in V, the phones were classified into 16 phones groups using KNN method (k=7).

Table 2 shows the results of phones period detections and phones recognitions for 12 different utterances of different males and females, and Table 3 shows confusion of 16 phones groups.

The average of phone period detection is 66.1% and the average of phones recognition is 51.3%.

VII. Obstacles to Upgrade Recognition

Some reasons for the existence of confusion in the phones recognition are:

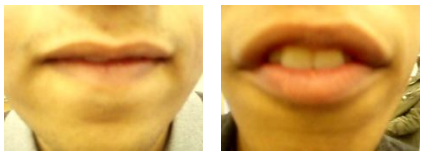
- Some phones of Table 1 are distinguished just by teeth whiteness, but the light condition and lips state of some people does not let this whiteness appear. For example, Fig. 13 shows the utterance starting frame of phone /fe/ uttered by two different speakers.
- In word recognition (e.g. [5], [11]), depending on the number of the syllables, several frames can be used to recognize the words. Although in this paper we tried to make the best use of the first and last frames of the phone utterance by computing the relative values, many differences between humans' lips cause the confusion of phones recognition. For example, Fig. 14 shows 6 different lips during the utterance of the same phone.
- Regarding the fact that this paper aims at recognizing the continuous phones in different words, the recognition of each phone depends on the lips state in the fore phone (as mentioned in IV-C). Although ignoring the first frames of each phone utterance can be useful but this problem is not eliminated completely. In addition, in some cases, the post- phone affects lips state and it causes confusion in phones recognition.

TABLE 2
Results of Phones Limit Detection and Phones Recognition

Subject Number	Phones Limit Detection [%]	Phones Recognition [%]
1	88.5	52.5
2	65.4	35.3
3	65	53.9
4	72	50
5	64.7	63.6
6	65	69.2
7	57.7	46.7
8	76.9	45
9	61.5	56.3
10	61.5	50
11	57.7	60
12	57.7	33.3

TABLE 3
The Confusion Matrix for 16 Phones Groups in Table 1

	G1	G2	G3	G4	G5	G6	G7	G8	G9	G10	G11	G12	G13	G14	G15	G16
G1	69.8	2.3	2.3	11.6	4.7	2.3	0	2.3	0	0	0	0	2.3	0	0	2.3
G2	25	50	0	0	0	25	0	0	0	0	0	0	0	0	0	0
G3	0	0	61.5	15.4	0	0	15.4	0	0	0	0	0	0	7.7	0	0
G4	7.7	7.7	0	30.8	0	0	7.7	15.4	0	0	0	7.7	0	0	0	7.7
G5	33.3	0	0	0	55.6	0	0	0	0	0	0	0	0	0	0	11.1
G6	0	0	0	0	0	100	0	0	0	0	0	0	0	0	0	0
G7	20	0	0	0	0	20	60	0	0	0	0	0	0	0	0	0
G8	0	0	7.7	15.4	15.4	7.7	15.4	23.1	0	0	7.7	0	0	0	0	7.7
G9	10	3.3	0	0	6.7	0	0	3.3	63.3	0	0	3.3	6.7	0	0	3.3
G10	0	0	0	0	20	20	20	0	0	40	0	0	0	0	0	0
G11	0	0	0	0	0	0	0	0	0	0	50	33.3	0	0	0	16.7
G12	0	0	12.5	1.5	0	0	0	0	0	0	25	25	0	0	0	25
G13	33.3	0	11.1	0	11.1	0	11.1	0	11.1	0	0	0	22.2	0	0	0
G14	9.1	36.4	0	0	0	0	0	0	0	9.1	0	9.1	9.1	27.3	0	0
G15	0	16.7	33.3	0	0	0	0	0	0	0	0	0	0	0	50	0
G16	0	0	10	0	0	0	10	0	0	0	0	30	10	0	0	40



(a) (b)

Fig. 13. Two different utterance starting frames of phone /fe/

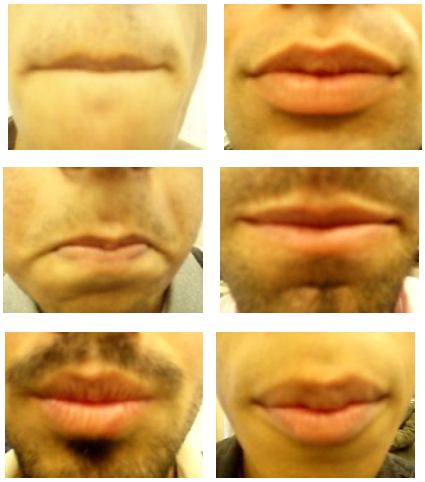


Fig. 14. 6 different lips during the utterance of the same phone

VIII. CONCLUSION

As we intended to move toward creating the real time and automatic lip reading system, in this paper we tried to recognize continuous phones in different words, from different persons. In addition, any pre- prepared hand-labeled model was not applied, and also for computing features values, we tried to process the least pixels to prevent a waste of time. The result of testing 12 different utterances was 66.1% for phones periods detection and 51.3% for phones recognition. We hope that someday a system will be created to decrease the obstacles to the relationship between deaf, dumb and other people.

REFERENCES

[1] R. Kaucic, B.Dalton and A. Blake, "Real-time lip tracking for audio-visual speech recognition applications," in *Proc. European Conf. Computer Vision, Cambridge, UK*, 1996, pp. 376–386.

[2] L. Matthews, S. Cox, R. Harvey and A. Bangham, "Lipreading using shape, shading and scale," in *Proc. Workshop on Audio Visual Speech Processing, Terrigal, Australia*, 1998, pp. 73–78.

[3] J. Luettin, N. A. Thacker and S. W. Beet, "Speaker identification by Lipreading," Proc. International Conference on Spoken Language Processing, 1996, pp. 62–65.

[4] T. Saitoh, K. Morishita and R. Konishi, "Analysis of efficient lip reading method for various languages," in *IEEE*, 2008, pp. 978-1-4244-2175-6.

[5] S. Nakamura, T. Kawamura and K. Sugahara, "Vowel recognition system by lip-reading method using active contour models and its hardware," in *SICE-ICASE International Joint Conference*, 2006, pp.1143-1146.

[6] G. I. Chiou and J. N. Hwang, "Lipreading by using snakes, principal component analysis, and hidden markov models to recognize color motion video," in *SICE-ICASE International Joint Conference*, 2006, pp. 18–21.

[7] T. Saitoh, M. Hisagi and R. Konishi, "Japanese phone recognition using lip image information," in *IAPR Conference on machine vision applications, Tokyo, Japan*, 2007, paper 03-27, pp. 134–137.

[8] S. Samadiyan, "Lip reading of limited set of Persian words," M. Eng. thesis, Teacher Training University, Tehran, Iran, Sep. 2005.

[9] R. Seguier and N. Cladel, "Genetic snakes: Application on lipreading," in *International Conference on Artificial Neural Networks and Genetic Algorithms (ICANNGA)*, 2003.

[10] B. J. Theobald, R. Harvey, S. J. cox, C. Lewis and G. P. Owen, "Lip-reading enhancement for law enforcement," in *SPIE Conference on Optic and Photonics for Counterterrorism and Crime Fighting, G. Owen and C. Lewis, Eds., vol.6402*, 2006, pp. 640 205–1–640 205–9.

[11] H. Soltani-Zadeh and S. Baradaran_Shokuhi, "Lip reading and visual speech recognition," in *ICME Conference on meca-tronic engineering, Islamic Azad University, Iran, 2003*, paper 93647, p. 182.

[12] A. Bagai, H. Gandhi, R. Goyal, M. Kohli and T. V. Prasad, "Lip-reading using neural networks," *IJCSNS International Journal of Computer Science and Network Security*, vol. 9 No.4, pp. 108–111, Apr. 2009.

[13] H. Soltani_Zadeh, "Speech recognition using visual tecniques," M. Eng. thesis, Science and Industry University, Tehran, Iran, Mar. 2002.