Visemes Recognition in Continuous Persian Words Using Lip Reading

Yousef Taheri
Department of Computer, Science and Research Branch, Islamic Azad University (IAU) Khouzestan, Iran
00989173055128
youtaheri@yahoo.com

Mohammad Hosein Yektaie
Department of Computer, Science and Research Branch, Islamic Azad University (IAU) Khouzestan, Iran
00989122192186
mh_yektaie@yahoo.com

Amir Masoud Rahmani
Department of Computer, Science and Research Branch, Islamic Azad University (IAU) Khouzestan, Iran
00989121784956
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 tried to recognize visemes (visual phonemes) in different continuous 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 it is the time to decrease the researches about lip reading systems that have different limitations and just increase recognition percentage from 99.1% to 99.3% or from 99.8% to 100% and so on. The result of testing 10 different utterances was 78.3% for locating visemes in continuous uttered words and 67.1% for visemes recognition.

Keywords

Lip reading, viseme, Continuous words, K-NN, Persian.

1. INTRODUCTION

To more efficiently make use of the lips, these interesting God's gifts to humans, we laid stress on leading our study in the field of lip reading by deaf and dumb people in communications. Because of almost high similarity between visemes in all languages and knowing that visemes are the smallest parts of the uttered words in speech reading, we focused on recognizing visemes. Due to the similarities of some visemes utterances created by lips and teeth, we classified all Persian visemes into 10 groups (6 vowels and 4 consonants), and we tried to recognize continuous visemes in different words uttered with different speeds and by different speakers, including males and females.

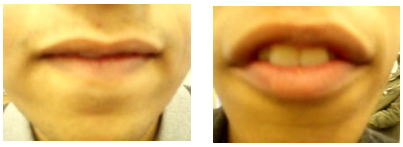
We neither carried out any steps of lip reading process by hand (e.g. [2]–[4], [10]), recognized separated phones or words (e.g. [4]–[6]), used speaker dependant method [7], nor used just one or a few subjects to create or learn our lip reading system (e.g. [8], [9]), because we believe that considering these specifications together in researches results in creating a real time automatic lip reading system in less time in future. In addition, we restrained the processing of the pixels that have a little effect on accuracy to increase the time of lip reading process, and therefore we can use the applied methods in this paper in a small instrument, even in mobile phones.

The reason for selecting the field of visual-only lip reading is to try using visual features as much as possible and hence more efficiently in comparison to other audio-visual lip reading systems (e.g. [1]), and then in future works, as visual features have a little role in speech reading, we'll try to use and add audio features to them for speech reading.

2. OBSTACLES TO UPGRADE RECOGNITION

Some obstacles that don't let our system do its best recognition are:

- Some visemes are distinguished just by teeth whiteness, but the light condition and lips state of some people does not let this whiteness appear. For example, Figure 1 shows the utterance frame of viseme /f/ uttered by two different speakers.



(a) (b)

Figure 1. Two different utterance frames of viseme /f/

- 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 single frame of the viseme by computing the relative values, many differences between humans' lips cause the confusion of visemes recognition. For example, Figure 2 shows 6 different lips during the utterance of the same viseme.

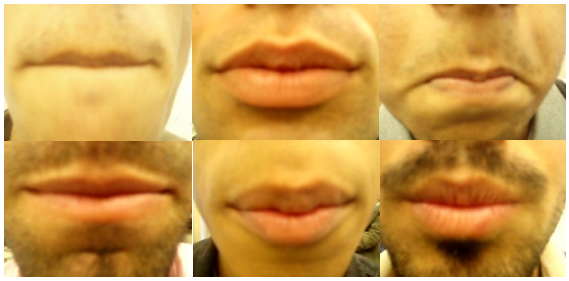
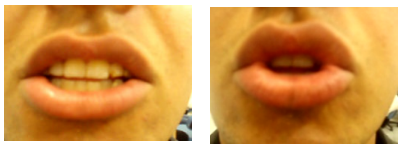


Figure 2. 6 different lips during the utterance of the same viseme

- Figure 3 shows the utterance frame of consonant /d/ belongs to one person, but before Figure 3(a), he uttered vowel /æ/, while before Figure 3(b), he uttered vowel

/əʊ/, and although both consonants are one, the lips states are different in these images. Regarding the fact that this paper aims at recognizing the visemes in different continuous words, the recognition of each consonant depends on the lips state in the pre- viseme. In addition, in some cases, the post- viseme affects lips state and it causes confusion in visemes recognition.



(a) (b)

Figure 3. Two different utterance frames of viseme /d/

3. 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. Therefore, in this paper we just locate 5 points on lips as shown in Figure 4.

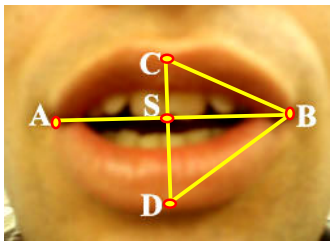


Figure 4. Locating 5 points on lips

To locate these points, a combination of some simple operations has been exerted on images:

- Converting to HSV format
- Converting to black and white images with appropriate thresholds
- Finding boundaries of black and white images
- And the instruction:

  Adjust(lips,[rl,rh],[gl,gh],[bl,bh]) (1)

  that 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.

4. FEATURE EXTRACTION

4.1. Proposed Features

To take the most advantage of points obtained in 3, we defined 8 following features:

1) $\frac{\text{The whiteness of region CSB}}{\text{Region CSB}}$

2) $\frac{\text{The whiteness of region DSB}}{\text{Region DSB}}$

3) $\frac{\text{The cavity appeared in region CSB}}{\text{Region CSB}}$

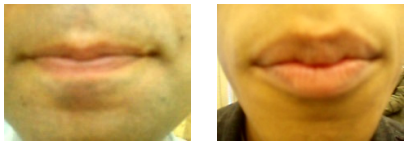
4) $\frac{\text{The cavity appeared in region DSB}}{\text{Region DSB}}$

5) $\frac{SB}{SC}$ 6) $\frac{SB}{SD}$

7) $\frac{SB}{SC+SD}$ 8) $\frac{SC}{SD}$

4.2. Using Criterion Frame

Figure 5 shows the utterance frame of viseme /b/ by different persons.



(a) (b)

Figure 5. Utterance frame of viseme /b/ by different persons

The value of feature 7, that represents the height to width ratio of lips, is much smaller in Figure 5(a) than this feature value in Figure 5(b), while this feature must have almost the same value for the same visemes when uttered by different persons. 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 the difference between the features values of the criterion frame (CFi) and the features values (Fi) :

$$F_i = CF_i - F_i \qquad (2)$$

F_i indicates i'th feature value and CF_i indicates the feature value of the criterion frame.

Hence, the thickness and thinness of lips can not be inconvenient in the visemes recognition.

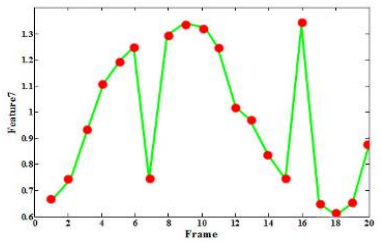
5. LOCATING VISEMES

After processing the images to extract the features, we processed the signals drawn by frame features in order to locate the visemes in uttered continuous words.

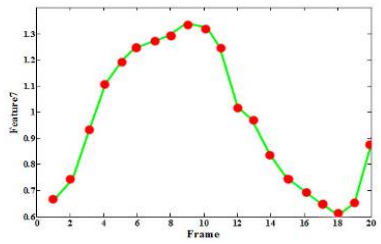
5.1. Normalization

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. For example, as shown in Figure 6(a), in spite of descending the signal, feature7 increased just for one frame, or in spite of ascending the signal, feature7 decreased just for one frame.

To delete the unsuitable effect of those frames on signal of feature7, we replaced the features values of such frames by features average values of their pre- and post- frames (as shown in Figure 6(b)). Thus, the bad effects of the wrong points located on lips in 3, are avoided to some extent.



(a)



(b)

Figure 6. part of feature7 signal for a word (a) and normalized signal of it (b)

5.2. Locating Visemes in Continuous Words

While the target of this paper was the recognition of visemes in continuous words, with different uttering speed, we tried to locate the visemes automatically by using features signals. After processing the signals and their various combinations, we concluded that feature7 can be helpful for detecting the frames of the vowels and consonants, because local maximum points indicate the number of the consonant frames and local minimum points indicate the number of the vowel frames. Hence, in Figure 7 that shows the signal of feature7 for uttering some Persian words, the local maximum and minimum points of the signal show the number of the visemes frames.

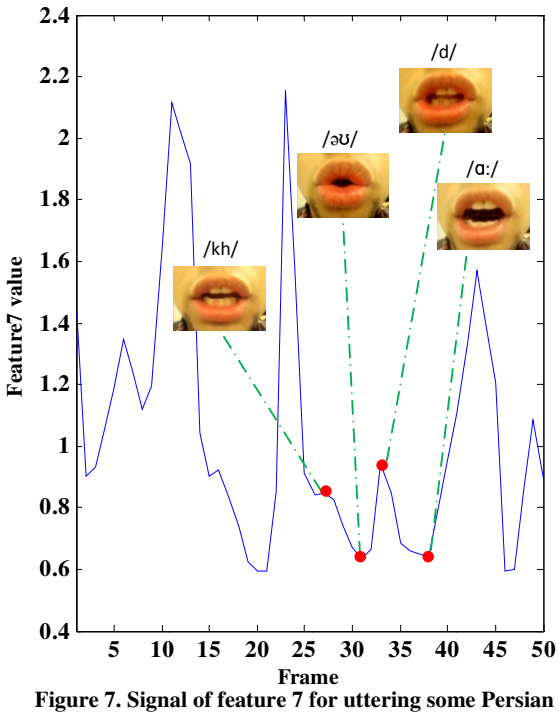


Figure 7. Signal of feature 7 for uttering some Persian words including 'khoda'

5.3. Reducing Pre- Viseme Effect

As mentioned in 2, the recognition of each consonant depends on the lips state in the pre- viseme; therefore, to reduce the pre-vowel effect on consonants, we overlooked one-fourth of the starting frames between consonants and post- vowels.

6. 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 4.2), the extracted features of individual viseme 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 visemes in this paper, they can not be useful, because unlike the words, for the visemes, features values do not increase and decrease through the utterance, and it is just a single state that can be useful.

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. We used the distance:

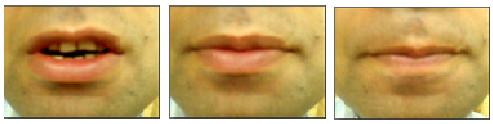
$$d^2_{test,train(i)} = \sum_{k=1}^{fn} \Big(\big(3 - \text{Symm}\big(\text{test(k)} * \text{train(i,k)}\big)\big) * |\text{test(k)} - \text{train(i,k)}| + 1 \Big)^2$$

$$where \quad Symm(\text{x}) = \begin{cases} 0, & x < 0 \\ 2, & x \geq 0 \end{cases} \qquad (3)$$

in KNN method, and as compared with the Euclidian distance:

$$d^2_{test,train(i)} = \sum_{k=1}^{fn} \big(\text{test(k)} - \text{train(i,k)}\big)^2 \qquad (4)$$

It has two advantages if it is used in lip reading system. In these relations, the vector 'test' includes the features values of viseme that should be recognized, the vector 'train(i)' includes the features values of the i'th trained viseme, and 'fn' indicates the number of the features. Figure 8 is sufficient to explain the first advantage of the proposed relation. Figure 8(a) shows the lips' state of the trained viseme /n/, Figure 8(c) shows viseme /b/ that should be recognized, and Figure 8(b) shows the criterion frame of this speaker.



(a) (b) (c)

Figure 8. Lips' states for uttering the visemes /n/(a) , /b/(b) and the criterion frame (b)

As shown in Figure 8, the difference between feature7 in Figure 8(a) and Figure 8(b) is almost equal to the difference between this feature in Figure 8(b) and Figure 8(c), while these two values must be different for visemes /b/ and /n/.Thus, we've added:

$$\dots\left(3-\mathrm{Symm}\left(\mathrm{test(k)}*\mathrm{train(i,k)}\right)\right)*\dots$$

$$where \quad Symm(\mathrm{x})=\begin{cases}0, & x<0\\ 2, & x\geq 0\end{cases}$$

to the Euclidian distance. By this addition, we've differentiated between such visemes, because the distance increases 3 times when the lips' states of the two visemes are different in comparison with the criterion lips' state.

Also, as another advantage of the proposed distance, we've used:

$$\dots|\mathrm{test(k)}-\mathrm{train(i,k)}|+1)^2$$

instead of:

$$\left(\mathrm{test(k)}-\mathrm{train(i,k)}\right)^2$$

because the value of $\mathrm{test(k)}-\mathrm{train(i,k)}$, can be smaller or greater than one for different features, while the square of a number between zero and one, unlike the other numbers, is smaller than that number. And this situation is not suitable for calculating the distance in KNN method that should be used in a lip reading system. Therefore, by adding '+1' to this value, all values (and subsequently their squares) become greater than or equal to one.

7. EXPERIMENT

Due to the similarities of some viseme utterances in Persian language, we classified all visemes into 10 groups (6 vowels and 4 consonants). In other words, homophonous specification was considered in this classification. Figure 9 shows the visemes that are classified in these groups.

1	/q/ /t/	/h/ /kh/	/d/ /r/	/gh/ /k/	/g/ /l/	/n/ /j/						
2	/θ/ /dʒ/	/tʃ/ /ð/	/z/ /ʒ/	/s/ /ʃ/			3	/b/ /p/	/m/			
4	/f/	/v/			5	/ɑ:/			6	/u:/		
7	/əʊ/		8	/i:/			9	/e/			10	/æ/

Figure 9. Classification of visemes into 10 groups

To obtain training sets, we used 7 word sets that are uttered by different people (males and females). These sets include visemes from all 10 groups. Then we extracted different visemes from the continuously uttered words as training sets. To test our system, different males and females uttered different continuous words with different speeds 10 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 located the visemes in uttered continuous words (as explained in 5). Then, as mentioned in 6, the visemes were classified into 10 groups using KNN method (k=7). Table 1 shows the results of visemes detections and visemes recognitions for 10 different utterances of different males and females.

8. CONCLUSION

As we intended to move toward creating the real time and automatic lip reading system, in this paper we tried to recognize the visemes in different continuous 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 10 different utterances was 78.3% for visemes detection and 67.1% for visemes recognition. We hope that someday a system will be created to decrease the obstacles to the relationship between deaf, dumb and other people.

Table 1. Results of visemes detection and recognition

Subject Number	Visemes Detection[%]	Visemes Recognition [%]
1	90.4	63.8
2	67.6	91.3
3	80.8	54.8
4	75	56.4
5	79.2	73.7
6	85	64.7
7	71.2	62.2
8	75	73.3
9	76.9	65
10	82	65.9
Mean	78.3	67.1

9. 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.