Vowel Recognition using Neural Networks

Vahideh Sadat Sadeghi†, Khashayar Yaghmaie††

†Faculty of Engineering, Semnan University, Semnan, Iran
††Faculty of Engineering, Semnan University, Semnan, Iran

Summary
Speech recognition techniques have been developed dramatically in recent years. Nevertheless, errors caused by environmental noise are still a serious problem in recognition. Employing algorithms to detect and follow the motion of lips have been widely used to improve the performance of speech recognition algorithms. This paper presents a novel technique to recognize vowels. Lip features extracted by using a combined method are used as input parameters to a neural network system for recognition. Accuracy of the proposed method is verified by using it to recognize 6 main Farsi vowels.

Key words:
Vowel recognition, visual features, neural networks

1. Introduction

Because of the existence of noise in many of circumstances, supplemental use of visual features such as mouth width and height is expected to improve the performance of speech recognition algorithms[1][2]. Automatic lip reading, however, is difficult for both the visual feature extraction and the speech recognition processes. Visual feature extraction requires a robust method of tracking the speaker's lips through a sequence of images and a representation of the inner mouth appearance. Lip tracking is not a trivial task, since there is variety in people in skin color, lip color, lip width, and the amount of lip movement during speech, as well as variability in the environment such as lighting conditions. Moreover, any method used to track lips during speech should not only be adaptive to the movement of the lips from frame to frame, but also stable enough not to be affected by the appearance of the teeth and tongue [3]. Regarding the recognition process, different methods have been developed to recognize speech according to the audio and visual features. For example the neural networks and hidden Markov models have widely been used in many problems such as classification and speech recognition [4],[5],[6].

This paper presents a method for using lip features in recognizing vowels. Techniques for extracting lip features are described in section 2. In section 3 a neural network is introduced to classify vowels. The proposed algorithm is evaluated by employing it in recognition of 6 main Persian vowels in section 4.

Manuscript received December 5, 2006.
Manuscript revised December 25, 2006.

2. Feature extraction

In the first stage, appropriate features of lips should be extracted. There are two major methods to extract the visual features called" Up to down" and "Down to up" methods. In the following, three down to up methods are introduced, based on which a method is introduced to extract visual features.

2.1 Binary transformation according to the luminance of the lip images

The simplest method to segment an image is to use a threshold value to group its pixels into black and white. Fig. 1 shows the effect of this method on the lip images.

Fig. 1 Binary transformation on the lip image

2.2 K-means algorithm

K-means clustering can best be described as a partitioning method [7][8] . That is the K-means algorithm partitions the data into k mutually exclusive clusters. In fact, K-means treats each observation in the data set as an object having a location in space. It finds a partition in which objects within each cluster are as close to each other as possible and as far as from objects in other clusters. In [9] a method for using K-means algorithm to extract suitable parameters for lip reading is introduced. An example for the application of this method has been shown in Fig. 2

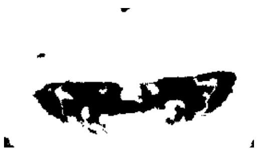


Fig. 2 Lip extraction using K-means algorithm

2.3 Red Exclusion Algorithm

In this technique green and blue levels of the image pixels are used to exclude lip from the other parts [10]. This method is based on the fact that face parts, including lips, are predominantly red, such that any contrast that may develop would be found in blue or green colors range, and thus the technique named red exclusion. It says that pixels belonging to lips have green and blue levels G and B so that

$$\ell og(\frac{G}{B}) \leq \beta \qquad (1)$$

In [8], however, no exact or approximate value for β is suggested. Fig. 3 shows the result of the red exclusion method applied to our image, using .6 for β. This value was found by performing extensive experiments on different images taken under different circumstances.

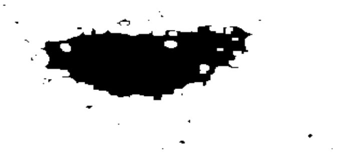


Fig. 3 lip extraction using red exclusion algorithm

3. A combined method for lip extraction

The described methods are able to extract lips fairly accurately. However, their performance depends highly on the subject of the image as well as the circumstances under which the image is recorded. In order to increase the accuracy of these methods a combined method is developed in which the results obtained by applying these techniques are combined to yield a higher accuracy algorithm. Fig. 4 shows the proposed method.

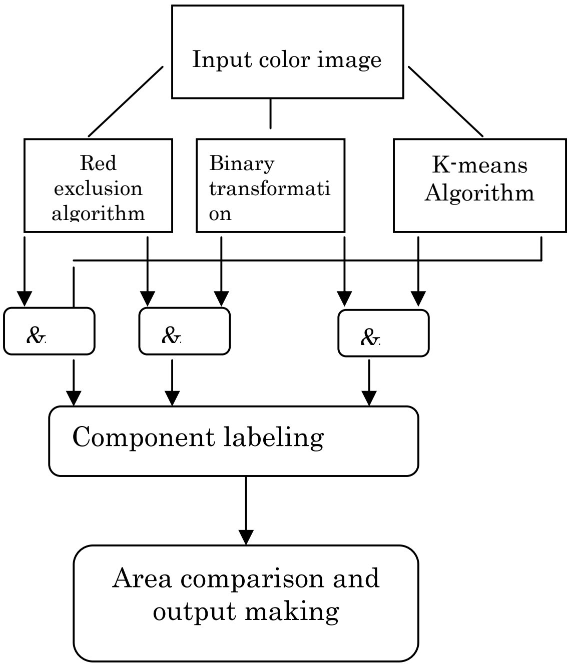
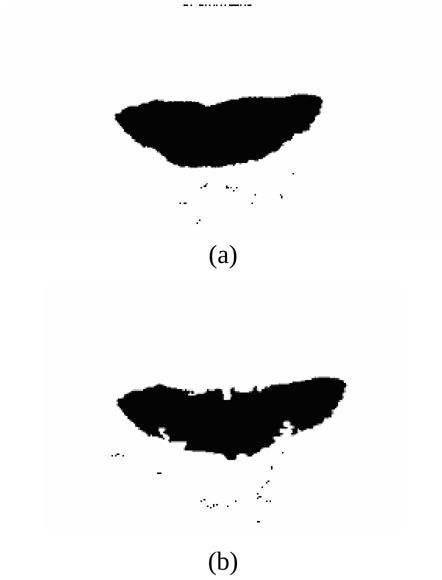


Fig. 4 The proposed algorithm for lip feature extraction

At the first stage the input image is processed the three main algorithms separately. The obtained results are then combined using the " &" operator (Fig. 5). In the third stage component labeling [11] is exploited to separate individual parts from each other and so to eliminate the excessive and noisy parts as depicted in Fig. 6.

Fig. 5 shows the result of the application of the proposed method to the image used in earlier experiments in this paper.



(c)

Fig. 5 (a) the effect of k-means and binary transformation , (b) the effect of red exclusion and binary transformation and (c) the effect of k-means and red exclusion

Fig. 6 Extraction of the largest part of first image in Fig.5 using component labeling

The above algorithm is used to extract the lip region. In the next stage some points on the recognized lip are marked so that other features such as mouth width (mouth opening in the horizontal direction) and height (opening in the vertical direction) can be extracted. Fig. 7 shows the points on the lips marked by the employed algorithm.

Fig. 7 Determination of some important points on the lips

4. VOWEL RECOGNITION

4.1. Vowels

Vowels and consonants are the basic elements of each language. Moreover, the difference in pronunciation of a word uttered by people of different mother tongues is mainly due to the variations in vowels and the way they are pronounced[12][13]. In Farsi (Persian mother tongue) there are 6 distinct vowels demonstrated as , , ,آ ,ایـ ,او which are fairly similar to English vowels a ,e ,o ,â ,i and u respectively. In this section a method is presented for vowel recognition using the lip features extracted in the former sections and applying them to an appropriate neural network. The developed method is then used to recognize the 6 major Farsi vowels.

4.2 The database

In order to prepare the suitable database 8 persons, (2 males and 6 females) were asked to utter 42 monosyllabic Persian words in which, each word contained a single vowel. The scenes of the utterances were then recorded by the common rate of 30 frames per second.

4.3 Implementation of the recognition algorithm on neural network

The neural networks have been widely used in many different problems such as speech classification and recognition [14][15]. Extensive experiments showed that a two layer neural network with 10 input, 25 hidden neurons and 6 outputs is suitable for recognition of the 6 classes of Farsi vowels. Normalized width (mouth opening in horizontal direction) and height (mouth opening in vertical direction) of the lips in 10 middle frames of the uttered words were used as the inputs to the neural network (NN). To exploit fast convergence and minimal storage the resilient back propagation training method with algorithms shown in Fig.8 and Fig.9 were used [16]. These networks are similar except that their activation functions in the hidden layer are different The decision algorithm is shown in Fig. 10. The two NNs in this figure are similar i.e. they are either of the networks introduced in Fig. 8 and 9. The decision algorithm is executed once using each network. In each run, the vowel giving the maximum output is selected as the candidate vowel for the exploited NN. In the final stage, as shown in Fig. 11, the two vowel candidates are compared and the one which has resulted in the maximum output is selected as the recognized vowel.

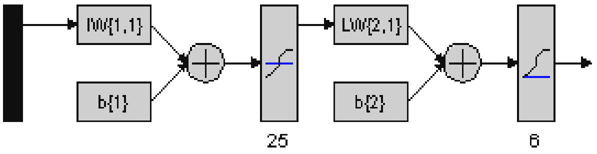


Fig. 8 Neural network with tan-sigmoid activation function in hidden layer and log-sigmoid activation function in output layer

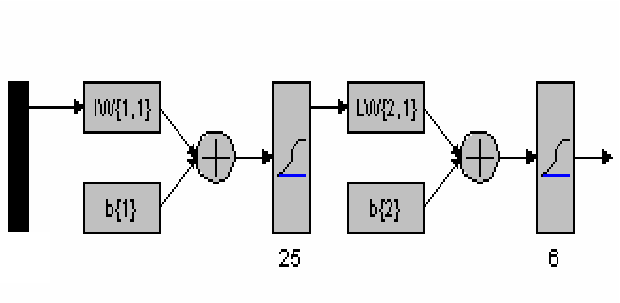


Fig. 9 Neural network with log-sigmoid activation function in hidden layer and log-sigmoid activation function in output layer

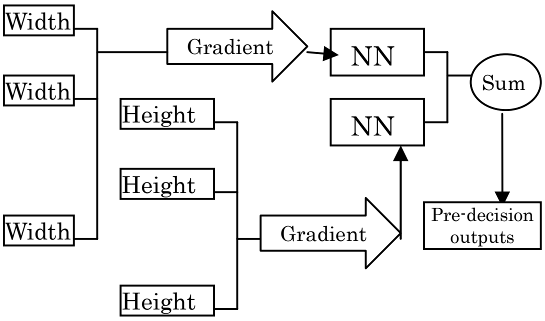


Fig. 10 Using the sum of network outputs as pre-decision

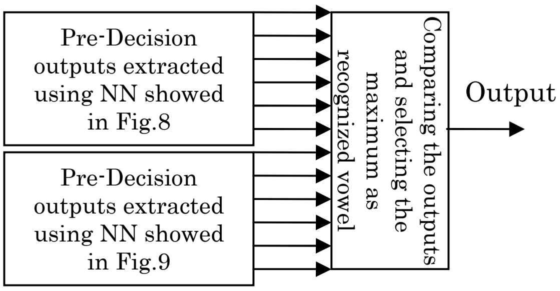

Pre-Decision outputs extracted using NN showed in Fig.8
Pre-Decision outputs extracted using NN showed in Fig.9
Comparing the outputs and selecting the maximum as recognized vowel
Output

Fig. 11 The final stage for vowel recognition

Table1. Recognition matrix using the maximum of pre-decision outputs as the final decision (in 100 utterance)

Recognized Uttered				آ	ايـ	او
	81	0	13	0	6	0
	10	42	0	12	25	11
	5	12	71	0	0	12
آ	4	0	0	96	0	0
ايـ	0	13	25	12	38	12
او	5	0	0	0	0	95

Noticing the speed of the employed algorithm, the accuracy of the results is promising and the low accuracy in recognition of and ايـ which are due to low impact of the model parameters by these vowels can be modified using similar parameters.

Meanwhile it is obvious from Table.1 that the total result in this way is 70.5% that is comparable with other similar works in other languages, for instance 70% in [6] that classifies 5 Japanese vowels uttered by 2 persons. Although our database was larger and our vowels was further .Since there was no similar work in Persian we couldn’t compare our results with them ,however it is also comparable with the result introduced in[17] that visually classifies 6 Farsi words uttered by 8 persons and is equal to 64.4%.

5. Conclusion

In this paper a simple and efficient method for extraction of visual features of lips was introduced. The results were then used as input data to a neural network for recognition of vowels in Farsi language. Despite the simplicity of the employed methods the early results are promising and other methods for improving the results are under development.

Acknowledgement

The authors would like to thank Iran Telecommunication Research Center (ITRC) for their support and help.

References

[1] C.Neti,Joint Processing of Audio and Visual Information for SpeechRecognition,www.clsp.jhu.edu/ws2000 /presentations / preliminarychlapathy_netineti_presentation.pdf/ reviewed on 10.05.2006

[2] B.P Yuhas.et.al.,Integration of acoustic and visual speech signals using neural networks. ,IEEE Communication Magazine November 1989

[3] M. Barnard , E. J. Holden, R. Owens," Lip tracking using pattern matching snakes,",5th Conf. on Computer vision,pp.1-6, 2002

[4] A.Waibel. Modular construction of time-delay neural networks for speech recognition, Neural Computation, 1,39-46 1989

[5] S.Nakamura, Statistical Multimodal Integration for Audio–Visual Speech Processing. IEEE transaction on neural networks, 13(4), JULY 2002

[6] T. Shinchi,et.al,Vowel recognition according to lip shapes using neural networks,. Proc. of IEEE 1998

[7] H.Spath,Cluster Dissection and Analysis: Theory, FORTRAN Programs, Examples, translated by J. Goldschmidt, Halsted Press,pp.226, 1985.

[8] McQueen .Some Methods for Classification and Analysis of Multivariate Observations," Proc. 5th Berkeley Symp. On Math. Stat. and Prob,1, 281-296,1967

[9] M.T.,Sadeghi, M.S. Tabatabayi,Fusion in Decision Making Stage for Lip Segmentation in Human Faces."Proc.14th conference on electrical engineering, Tehran.Amirkabir University of technology. (in Persian)

[10] T.Lewis,D. Powers. Lip feature Extraction Using Red Exclusion, www.cs.usyd.edu.au/~vip2000 , reviewed on 2006.09.02

[11] R. M., Haralick, GL. Shapiro, Computer and Robot Vision, Volume I, Addison-Wesley, , 28-48,1992

[12] W. rui et.al. , Recognition of sequence lip images and application, Proc. ICSP 1998

[13] T. E.TobeIyt,et.al., On-Line Speech-Reading System for Japanese Language, 2000

[14] B.P. Yuhas, et.al, Neural network models of sensory integration for improved vowel recognition ,Proc.IEEE 78(10),1658-1668, 1988

[15] P. Teissier,et.al.Comparing Models for Audiovisual Fusionin a Noisy-Vowel Recognition Task, IEEE transaction on speech and audio processing , 7(6), November 1999

[16] V.S.Sadeghi.Vowel Recognition in Persian Monosyllabic and Disyllabic Words,Msc.project,University of Semnan, Iran,101-105 2006

Vahideh Sadat Sadeghi was born in 1982 in Khansar,Iran. She received his B.S.c in electronics from Isfahan University of technology; Iran in 2003.She is currently a M.Sc student in electronics, University of Semnan and working on lipreading as her thesis. She’s interests include image enhancement and image processing.

Khashayar Yaghmaie was born in 1957 in Semnan, Iran. He received his M.Sc. in telecommunication from Tehran University, Iran, in 1985 and his Ph.D in speech processing from university of Surrey, U.K. He is currently a lecture at the electrical engineering department of Semnan University. He’s interests include voice and speech coding and image processing.