ISSN (Print): 2320-9798
ISSN(Online): 2320-9801

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 1, Issue 9, November 2013

LIP MOTION SYNTHESIS USING PRINCIPAL COMPONENT ANALYSIS

Disha George[1], Yogesh Rathore[2]

P.G. Student, Department of Computer Science & Engineering, RITEE, Raipur, Chhattisgarh, India [1]

Sr.Lecturer, Department of Computer Science & Engineering, RITEE, Raipur, Chhattisgarh, India [2]

Abstract: Current studies states that not only audio but also video signs are delivering information on speech recognition. This feature can be used as a supplementary in the field of animation and lip motion reading for the enhancement of speech recognition. It has gained a wide attention in audio-visual speech recognition (AVSR) due to its potential applications. This research is divided into two-phases: (i) Firstly, taking frames and extracting features to be kept in database as standard. (ii) Secondly, having test image samples to be trained in neural network to check the alphabet spoken by recognizing what the person has spoke. Lip reading system has been developed using Principal Component Analysis using input images and 60% success has been achieved in the test phase in similar alphabets lip movements (such as u, o, q, b, e, i, l, n etc.).

Keywords: Lip reading, Eigen values and Eigen vectors, Principal Component Analysis, Neural Network.

I. INTRODUCTION

Lip-reading has been practised over many years to teach dumb and deaf pupils or persons to recognize what the front person has spoken and communicate in a effective way with other people[1].Extracting the audio bits of information of the speaker is itself a tedious task and brings complexity whereas visual signs also play a major role as supplementary to convey information as what does speaker says[2].Lip-reading is generally examined in two context: that is speech recognition and analysis of visual signs[8].Systems relied on audio waves are not preferable since signs are affected by different noises drastically. It is found that systems using both audio and video information are best option for healthy communication [2].Visual information constitutes 1/3 of the conveyed message [5, 6, 7].Some audio can be mixed with the environmental noises and they can be differentiated in visual space [4].In this research, videos of different subject has been taken and from which frames has been selected to conduct processing. The video corpus is converted into images of different alphabets over which the further work is performed. These images are selected manually by visualizing the changes in the frames during transitions [1].According to the work; there will be difference in the lip movements pronouncing different alphabets or letters of English even in similar sounding letters. This will include various parameters such as height between the inner lips, outer lips and similarly width between the inner lips[2].The vertical and horizontal distance between the lips vary considering the close approximations between similar pronouncing letters[1].Based on this research, creation of the database of commonly used alphabets is done and our neural network can be trained to find the best match between the input images and test sample images to find the letter with the closest proximity by its intelligent approach. The feature extraction is done from lip shape to form feature vector by Principal Component Analysis (PCA) to get the feature points in the database of various subject of speaking. A database is generated having feature points of various subjects. This database is compared with the test images to find the closest proximity among them, so that the image can be depicted what the person has spoken. Here, Neural Network RBFN (Radial Basis Function Network) is used in training purpose to get the best matched lip shape with the spoken alphabet to be identified [10].

II. LIP READING SYTEM

Lip reading system constitutes an operation which helps to understand what speaker has spoken without the requirement of audio information. It comprises of complex computational processes. The proposed system is subdivided into sub modules and each sub module is processed and analysed separately to collect every bit of information deeply.

In the proposed model, firstly the set of images is taken and then the lip portion is focussed as a part of work to be processed. Later, points in the contour of the lips are gathered from the frame to form the feature vector matrix. A feature matrix is drawn out by realizing operations on lip shapes and feature vector extraction for the following frames of various subjects. This feature matrix is inserted into the database after training them with Principal Component Analysis (PCA) technique. When a test sample image is assessed, then the same procedure is applied and is compared with the database maintained of the frames. The feature extraction is done and this vector of new image is compared to database to find the close proximity with any of the image feature vector so as to find what the person says. Few steps of operation are applied can be seen further.

A thresholding is applied to determine the lip region. Binary lips are obtained after transformation by strengthening the lips. This helps to detect the lip area and give a particular shape of the determined lip shape. Feature Vector is obtained from the binary image. The most work to be done in lip-reading is the feature extraction of the supposed lip image. The properties to be worked over are *the height between the inner lips, the height of outer lips, visibility of teeth in between, the corners of the lips, width of edges of lip, the coordinates representing the contour of lips.* The feature vector is prepared by finding these all properties one by one. This can be shown in the figure below:-

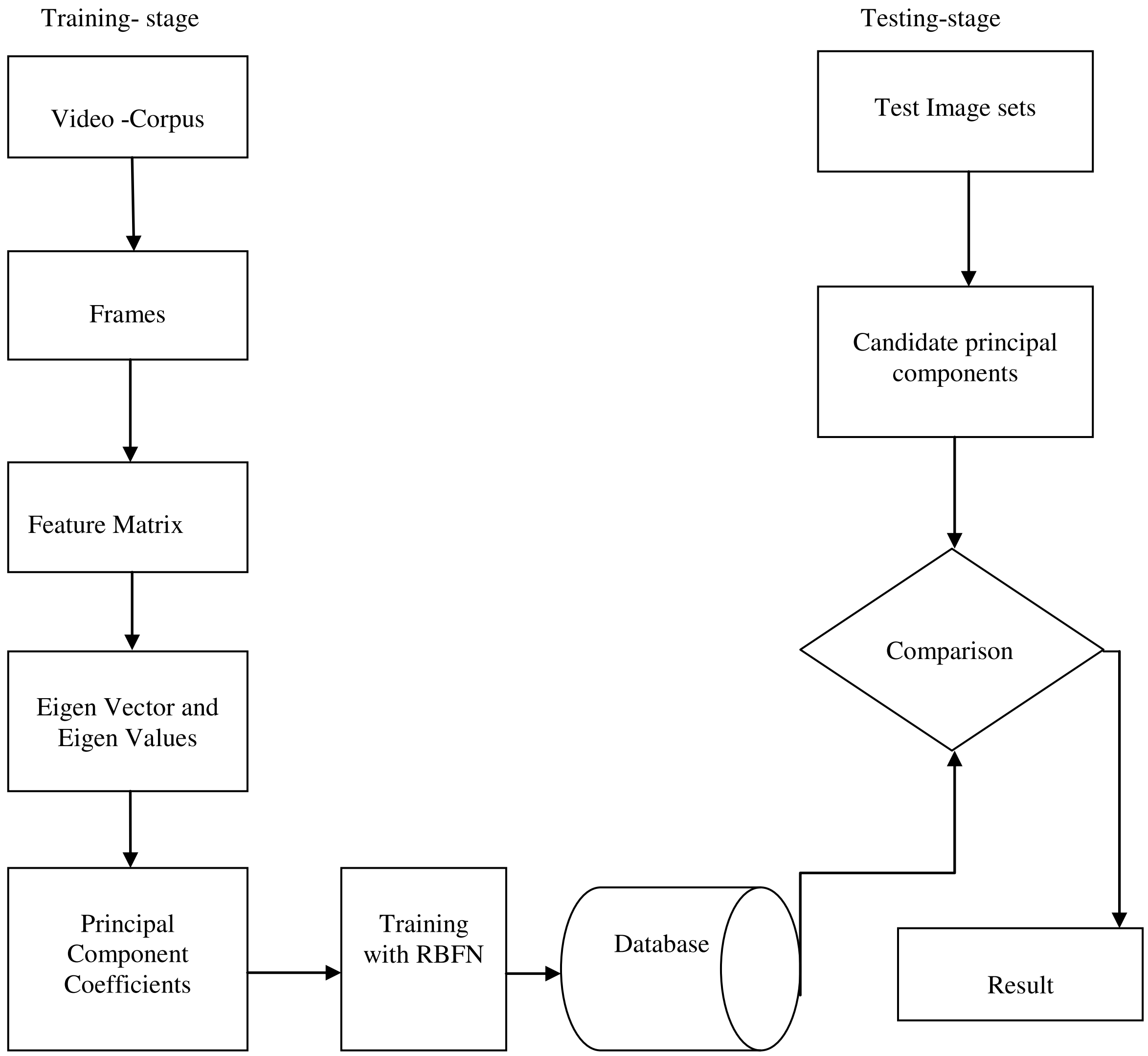

Fig.1. The block diagram of lip-reading system

A. *LIP READING USING PCA*

Principal Component Analysis is one of the best statistical methods by which face recognition and image compression aspects can be covered. It is a beneficial technique of finding patterns in high dimension data. It is a useful way of finding patterns so as to highlight the proximities and differences between the patterns[9].Once the pattern is found, its easy to compress the image[9].PCA method is comprised of two sub modules: in training phase and testing phase.

I) *Training- Stage:* The training phase is having recording of data by unsupervised learning. These image sets are limited to *n* frames to provide identical and uniformity. In this paper, the constant *n* is taken as 15. The feature vector consists of the properties stated above. In this phase, to apply feature vector to PCA method, it should be rearranged. Each row in the PCA feature matrix corresponds to a P_{fq} feature matrix. Thus, the training phase matrix is composed for PCA by altering the features matrix of each pattern to a row vector. The feature matrix for uth pattern and the feature matrix used for PCA is shown in Equation (1).

$$P_{fq_i} = [1_{f,}2_f, \dots .n_f]^T \qquad , \; PCA_{fq} = [P_{fq_1}, P_{fq_2}, \dots P_{fq_l}]^T \qquad (1)$$

where P_{fq_i} is the feature matrix of u^{th} pattern,PCA_{fq} is the matrix utilized in PCA and l is the pattern number utilized for training.Later,subtract the mean value of each row from each element in that row of PCA_{fq} to create the matrix *B* and the covariance of matrix *B* is calculated as seen in Equation (2).

$$\overline{X} = \frac{\sum_{i=i}^{m} X_i}{m}, \qquad B_k = X_{n_i} - \overline{X} \;, \qquad C = \frac{B^T.B}{(m-1)} \;, \qquad n = 1,2,\dots,l \qquad (2)$$

where *C* is the covariance matrix used. After that, the eigen values and eigen vectors of the *C* are calculated. The eigen values are arranged and then corresponding eigenvectors are sorted out according to the eigen values. The largest value of all the eigen values is the first component of the set. Lesser values of eigen values and the corresponding eigen vectors are eradicated, so that the dimension of the data set is reduced. The matrix comprising of the left eigen vectors is named transformation matrix *T*.

The primary aspect is obtained by multiplying the transformation matrix *T* and corresponding row of the of the feature matrix, because each row vector in the feature matrix formed by real values is associated with a pattern [9]. The relation of the process is given in Equation (3).

$$A_n = I_n \mathrm{x} U \qquad (3)$$

where *I* is a row vector obtained from transformation matrix with dimension *l*, *U* is a *l*x*m* transformation matrix and *A* is resulting vector. Additionally, *l* is the number of features, *m* is the new dimension resulted by reduction done on the phase of selection of principle components.

The last phase of PCA is storing primary aspects into the database. All of the aspects produced for each pattern are stored into the database in the phase of primary aspects generation. However, transformation matrix *U* used in the recognition is stored into the database also.

II) *Lip Reading-Stage:* In this stage, the identification of the data set which is trained before is applied. For this, firstly obtain the feature vector from test data set. This procedure (extract the feature vector) is similar with the training phase. Each row vector in the extracted feature vector is produced by transformation matrix U which is attained and stored into the database prior. Later, the candidate primary aspect is formed. The production of candidate primary aspect *A'* is derived from the equation $A' = I_n'$ x *U*. After the production of the candidate primary aspect, it is matched up with all the other aspects in the database. The most similar aspects are received as the same after match and these close proximities are accepted ones to be recognized as the lip shape of matched alphabet. This renders the most significant and matched alphabet to what the speaker said.

III.RADIAL BASIS FUNCTION NETWORK

A radial basis function network is a network similar to an artificial neural network. They accept numerical inputs, and produce some numerical outputs and can be used to make calculations. In training, RBFN is to discover the set of weights and bias values that build a network whose outputs best match with those of the training data. Training a radial basis function network comprises three important steps. In the primary step, a set of centroids is resoluted, a centroid for every hidden node. In the second step, a set of widths is resoluted, individual width value for every hidden node. In the third step, a set of weights is calculated. A set of properties, as mentioned earlier in the paper is used to train to find the similar pattern with the testing sample and if they matches, then that will be the best match and considered output. Radial Basis Function Network (RBFN) is an efficient solution for the researchers who are working on the field of machine recognition, pattern recognition and computer vision. The key challenge in the face recognition technology is to deliver high recognition rate. In this paper, a beneficial method has been presented for pattern recognition using principal component analysis and radial basis function. To be precise, principal component analysis has been used for feature extraction of the dataset and radial basis function network has been used as a classifier for classification of data and for recognition process also.

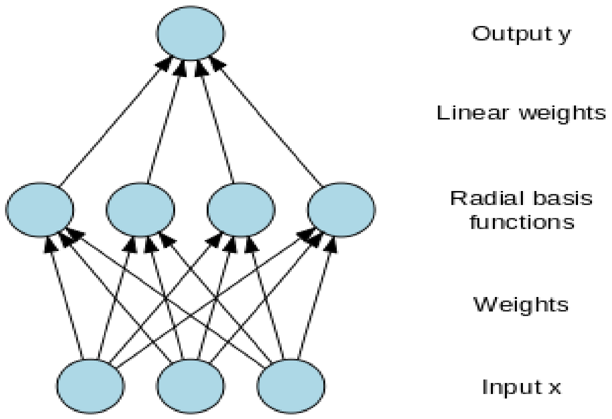

Fig. 2 Radial Basis Function network architecture

During Training stage , the input data set is given and the best matched features with the test mouth shape or the closest properties of the mouth shape which is similar to the input is the recognised letter. With the help of weight adjustment, this is the core of RBFN benefits to attain the goal; that is to get the desired output using radial function as well.

IV.RESULT AND CONCLUSION

In this paper, we have proposed a mechanism by which we can find the alphabet spoken recognising the lip shape of the speaker. Here, we have used the Radial Basis Function Network (RBFN) for the training and pattern recognition purpose. The Principal Component Analysis is used for the dimensionality reduction of high dimension dataset and to find the similarities both. The paper delivers better approximation between the patterns identified. In training stage, we undergone with some patterns.

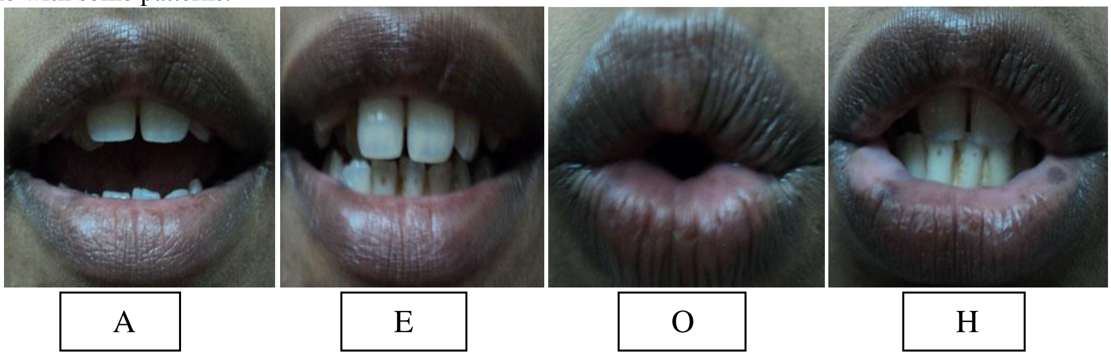

Fig. 3 Frames taken from videos of particular subjects

TABLE I
RESULTS OF SOME ALPHABETS

Alphabets	Recognition Results		
	First nearest	Second nearest	Third nearest
A	I (65%)	İ (20%)	A(70%)
E	E (30%)	I (70%)	B(65%)
B	B (80%)	E (25%)	I (30%)
H	H (80%)	J (10%)	G (5%)
O	U (60%)	I (20%)	Q (30%)
S	S (70%)	X (35%)	Z (10%)
U	U (80%)	I (60%)	O (60%)

In spite of this, we achieved about 60% of success in lip reading using PCA with computer. So, it can be concluded that if we take visual information auxiliary to audio information, we get more successful results in human-computer interaction. It can be demonstrated as:

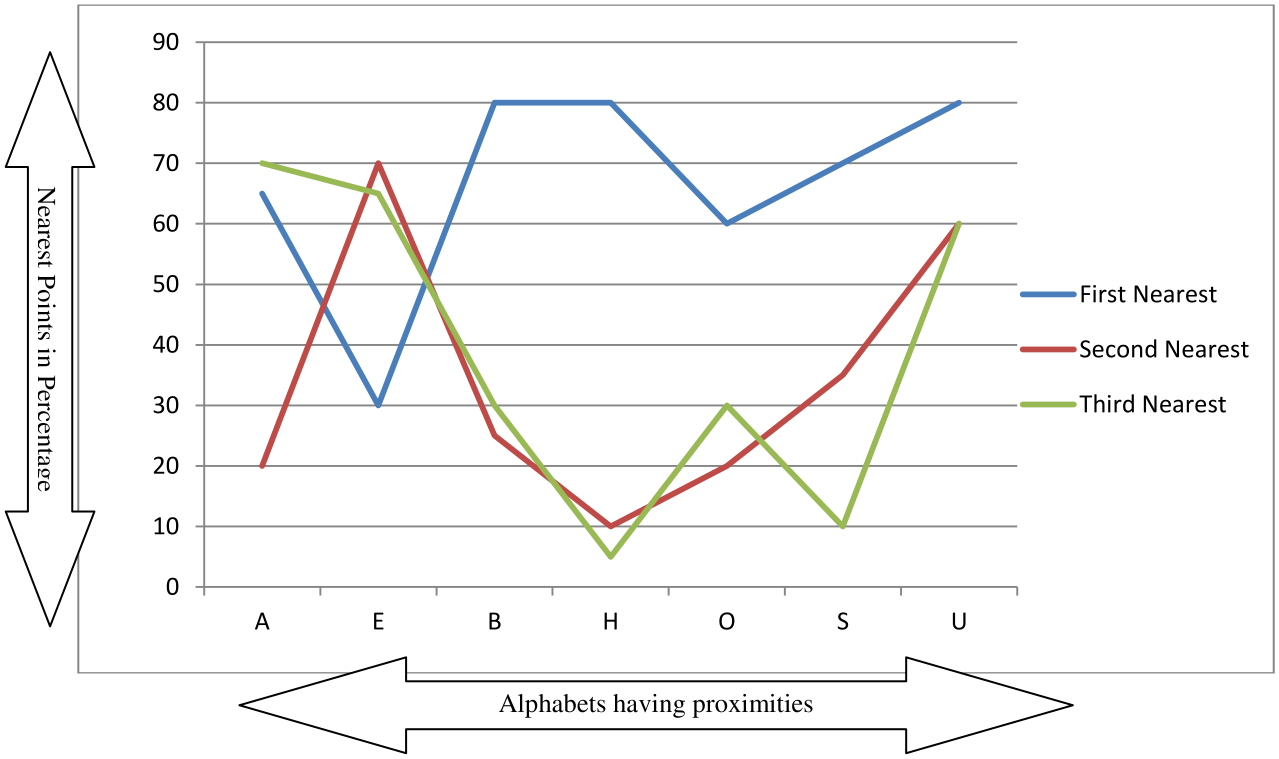

Fig. 3 A line graph between the closest similarities between the alphabets

The graph depicts the relation between alphabet proximities and the similarities among the alphabets falling nearby to the test sample of alphabet. It is seen from the graph that similarities can be found between the nearby lip shapes of various subjects but the alphabet having the most similarity will be the exact one.

REFERENCES

[1] Abhay Bagai, Harsh Gandhi, Rahul Goyal,Ms. Maitrei Kohli, Dr. T.V.Prasad,” Lip-Reading using Neural Networks” IJCSNS International Journal of Computer Science and Network 108 Security, VOL.9 No.4, April.

[2] Zafer Yavuz, Vasif Nabiyev ,"AUTOMATIC LIPREADING WITH PRINCIPLE COMPONENT ANALYSIS", supported by Karadeniz Technical University Research and development fund for the science project N.2005,112.009.01.

[3] V.V. Nabiyev, *Artificial Intelligence: Problems-Methods-Algorithms* (in Turkish), Seckin Publishing, 2nd Press, (2005).

[4] I. Matthews, T.Cootes, J.Bangham, S.Cox, R. Harvey, *Extraction of Visual Features for Lip reading*. PA&MI, IEEE Transaction, 198-213, (2002).

[5] A.Rogozan, P.Deléglise, Visible Speech Modeling and Hybrid Hidden Markov Models / Neural Networks Based Learning for Lipreading. IEEE Computer Society, France (1998).

[6] L.Xie, X.Cai, Z.Fu, R. Zhoa, D.Jiang. *A* Robust *Hierarchical Lip Tracking Approach for Lip reading and Audio Visual Speech Recognition.* Proceedings of the 3rd ICMLC, 3620-3624, Shangai, China, (2004).

[7] S.Wamg, H.Lau, S.Leng, H.Yan. *A Real Time Automatic Lip reading System.* ISCAS'04, vol:2, p.101-104, Hong Kong, (2004).

[8] Z.Yavuz, V.V.Nabiyev. *Automatic Lipreading*, 15th SIU'07, Eskişehir, (2007).

[9] I. S. Lindsay. A Tutorial on Principal Components Analysis, USA, (2002).

[10] Jiyong Ma, Member, IEEE, Ron Cole, Member, IEEE, Bryan Pellom, Member, IEEE,Wayne Ward, and Barbara Wise," Accurate Visible Speech Synthesis Based on Concatenating Variable Length Motion Capture Data", IEEE TRANSACTIONS ON VISUALIZATION AND COMPUTER GRAPHICS, VOL. 12, NO. 2, MARCH/APRIL 2006.

[11] R. Cole, S. Van Vuuren, B. Pellom, K. Hacioglu, J. Ma, J. Movellan,S. Schwartz, D. Wade-Stein, W. Ward, and J. Yan, "Perceptive Animated Interfaces: First Steps toward a New Paradigm for Human-Computer Interaction," Proc. IEEE, vol. 91, no. 9, pp. 1391-1405, 2003.

[12] G. Geiger, T. Ezzat, and T. Poggio, "Perceptual Evaluation of Video-Realistic Speech," CBCL Paper #224/AI Memo #2003-003, Mass. Inst. of Technology, Cambridge, Mass., Feb. 2003.

[13] S.W. Choi, D. Lee, J.H. Park, and I.B. Lee, "Nonlinear Regression Using RBFN with Linear Sub Models," Chemometrics and Intelligent Laboratory Systems, vol. 65, no. 2, pp. 191-208, 2003.

[14] J. Ma, J. Yan, and R. Cole, "CU Animate: Tools for Enabling Conversions with Animated Characters," Proc. Int'l Conf. Spoken Language Processing, pp. 197-200, 2002.

[15] J. Ma and R. Cole, "Animating Visible Speech and Facial Expressions," The Visual Computer, vol. 20, nos. 2-3, pp. 86-105, 2004.