Journal of Applied Sciences

ISSN 1812-5654

science alert

ANSInet
an open access publisher
http://ansinet.com

Journal of Applied Sciences 8 (2): 280-287, 2008
ISSN 1812-5654
© 2008 Asian Network for Scientific Information

A UMACE Filter Approach to Lipreading in Biometric Authentication System

Dzati Athiar Ramli, Salina Abdul Samad and Aini Hussain
Department of Electrical, Electronic and Systems Engineering, Faculty of Engineering,
Universiti Kebangsaan Malaysia, 43600 Bangi, Selangor, Malaysia

Abstract: Visual speech information, for example the appearance and the movement of lip during speech utterance can characterize a person's identity and therefore it can be used in personal authentication systems. In this study, we propose a novel approach by using lipreading data i.e., the sequence of entire region of mouth area produced during speech and the implementation of the Unconstrained Minimum Average Correlation Energy (UMACE) filter as a classifier for biometric authentication. The system performance is also enhanced by the implementation of multi sample fusion scheme using average operator. The results obtained from using a Digit Database shows that the use of lipreading information and UMACE filter has good potentials and is highly effective in reducing false acceptance and false rejection rates for speaker verification system performance.

Key words: Unconstrained Minimum Average Correlation Energy (UMACE), speaker verification, lip movement, biometrics

INTRODUCTION

Biometric authentication is the task of conforming or denying a person's claimed identity based on specific information either on his/her physiological or behavioral characteristics (Kittler *et al*., 1997). In speaker verification systems, primarily acoustic speech signals have been widely used for personal authentication systems (Furui, 1997; Campbell, 1997; Reynolds, 2002; Campbell *et al*., 2003). According to Reynolds (2002), the main strength of speaker verification technology using the speech signal to identify a person is that it is natural and unobtrusive to produce, requiring little custom hardware, has low computation requirement and is highly accurate.

However, the weaknesses are also observed by this modality. First, speech is a behavioral signal that may not be consistently reproduced by a speaker and can be affected by a speaker's health, for example if he has a cold. Second, varied microphones and channels can reduce accuracy. The biggest challenge is also discovered when the system is implemented in uncontrolled and harsh acoustic environment such as in cars and crowded airports. Therefore, many efforts have been done to cope with these setbacks (Campbell, 1997; Reynolds, 2002).

One solution to address these limitations is to combine the audio-based biometric system with visual-based biometric system. So that when the speech signal is somehow degraded, the other traits will still lead to an accurate decision. It is well-known that the visual modality of the speaker's mouth region provides additional speech information which can improve speaker recognition and verification system's performance (Wark *et al*., 1997, 1999; Broun *et al*., 2002). In general, visual features for lip verification can be classified into two main groups which are lip contour-based features and appearance-based features (Hennecke *et al*., 1996). For the lip contour based features, inner and outer lip contour and geometric features such as mouth height, width and angle are used as features to the verification system. Whereas, in appearance based group, the entire region containing the speaker's mouth is deemed as informative to authenticate the lip. Here, the Discrete Cosine Transform-based images are used as features to the system (Potamianos and Neti, 2000; Potamianar *et al*., 2003).

Many researchers based on lip information has been investigated recently. By extracting the shape and intensity information from speaker's lip, Wark and Sridharan (1998) succeed to develop a speaker verification system based on lip information features. Broun *et al.* (2002) worked on the extraction the geometric dimensions i.e., height, width and angle of the mouth while utters the words using MRF based lip segmentation. It was claimed to be highly effective in enhancing the accuracy of the speaker verification system. Apart from that, lip features based on Discrete Cosine Transform (DCT) and optical

Corresponding Author: Dzati Athiar Ramli, Department of Electrical, Electronic and Systems Engineering, Faculty of Engineering, Universiti Kebangsaan Malaysia, 43600 Bangi, Selangor, Malaysia
Tel: +603-89216331 Fax: +603-89296146

flow features of lip motion have been investigated by Fox and Reilly (2004) and Faraj and Bigun (2006), respectively.

In term of classifier for the verification task, approaches have spanned from simple template matching to dynamic time warping and recently to neural networks, Hidden Markov Model (HMM) and Gaussian Mixture Model (GMM) (Reynolds, 2002). A Support Vector Machine (SVM) for speaker recognition has been investigated by Campbell (2003).

In this study, an alternative technique for features as well as classifier is proposed. We aim to avoid the extraction of features such as height, width, angle and intensity as described in above reviews which is tedious and time consuming. We propose a novel approach by introducing the image of mouth region as features and the Unconstrained Minimum Average Correlation Energy (UMACE) filters are utilized for the verification task. We also suggest the implementation of multi-sample fusion technique to the system in order to enhance the reliably of the system performance. Poh *et al.* (2002) proved that multi sample fusion can improve the biometric system performance. Furthermore, the implementation of multi sample fusion approach does not give any burden to the system because utterance can always be split into many single frames.

The first objective of this study is to investigate the performance of our proposed speaker verification system and the second objective is to evaluate the significant of the multi-sample fusion technique. Our proposed features consist of a frame sequence of speaker mouth region (lipreading sequence images), captured from video while the speaker utters the speech. Our features contain physiological characteristic as well as behavioral characteristics as the changing of the appearance of the lip in each frame sequence provides extra information i.e., the way how people speak. So, this offers an advantage to the system instead of using static images.

The first motivation to implementing UMACE filter in this study is because UMACE filters use image as features for verification. So there is no tedious features extraction involved in this system in order to extract the lip features as in other classifiers. Second motivation is because of its benefits such as shift-invariance, ability to trade-off between discrimination and distortion tolerance i.e.,the variations in pose, illumination and facial expression. Consequently, implementing UMACE filters to this project is significant to deal with the changing of lip appearance in the lip sequence. Many researches on correlation filter for the purpose of biometric system have been done so far. Face verification and fingerprint verification using correlation filters have been investigated in Savvides *et al.* (2002) and Venkataramani and Vijaya Kumar (2003), respectively. The implementation of correlation filter for lower face verification centered on lips with different facial expression is also reported in Samad *et al.* (2007a). The study reported that, the results offer good potential compare with face verification performance and by using the bottom half of the face requires half the storage capacity of that of a full-face while giving comparable verification performance. In Samad *et al.* (2007b), we investigate the performance of lip motion sequence as features and UMACE filter is used as classifier for person identification system.

MATERIALS AND METHODS

Unconstrained Minimum Average Correlation Energy (UMACE) filters: The application of the correlation filter in biometric verification system can be found in Savvides *et al.* (2002) and Venkataramani and Vijaya Kumar (2003). The theory and the explanations of the design of the correlation filter are described in a tutorial survey paper (Vijaya Kumar, 1992). In earlier times, correlation filter was used for detecting a known reference image in the presence of additive white Gaussian noise. In order to be robust under variation of image i.e., scale, rotation and pose of the reference image, the Synthetic Discriminant Function (SDF) filter and the Equal Correlation Peak SDF (ECP SDF) filter are introduced. Here, several training images are allowed to be represented by a single correlation filter. By implementing these filters, the pre-specified peak value is produced when the testing image is tested. The peak corresponds to the authentic class or imposter class (Savvides and Vijaya Kumar, 2003).

The Minimum Average Correlation Energy (UMACE) filter variant and the Unconstrained Minimum Average Correlation Energy (UMACE) filter variant can be used in order to produce a sharp peak when the test image belongs to the authentic class. The MACE filter variant works by minimizing the average correlation energy of the training images and at the same time constraining the correlation output at the origin to a specific value. Whereas, the operation of the Unconstrained Minimum Average Correlation Energy (UMACE) filter minimizes the average correlation energy of the training images while maximizing the correlation output at the origin (Savvides and Vijaya Kumar, 2003). Both filters produce a similar result but UMACE filter is better than MACE filter in term of its simple computation. The equations of MACE and UMACE filter are shown as in Eq. 1 and 2, respectively.

$$H_{mace} = D^{-1}X\left(X^{+}D^{-1}X\right)^{-1}c \tag{1}$$

$$U_{mace} = D^{-1}m \tag{2}$$

D is a diagonal matrix with the average power spectrum of the training images placed along the diagonal elements. X consists of the Fourier transform of the training images lexicographically re-ordered and placed along each column. c is a column vector of length N containing the desired correlation output at the origin for each training images. Finally, m is a column vector containing the mean of the Fourier transforms of the training images (Savvides and Vijaya Kumar, 2003).

Figure 1 shows the verification process using correlation filter. By employing several training images for filter design, the test image is then cross-correlated with the template filter. The peak value produced by the correlation output determines the test image as an authentic or imposter class. Examples of the correlation plane for the test image from the authentic and imposter class are shown in Fig. 2 and 3, respectively.

The measurement of the sharpness of the peak is based on the Peak-to-Side lobe ratio (PSR) metric. The PSR calculates the largest peak yield from the correlation plane by calculating the mean and standard deviation from the 20×20 side lobe region with exclusion of the 5×5 central mask (Savvides and Vijaya Kumar, 2003). It is given as in Eq. 3 below.

$$\mathrm{PSR} = \frac{\text{peak - mean}}{s} \tag{3}$$

Multi sample fusion technique: In multi sample fusion, several samples of single biometric modality are employed to the system. According to Kittler *et al.* (1998) and Cheung *et al.* (2004), multi sample fusion is a successful technique to increase the performance of biometric system. Minimum, maximum, average, median, majority vote and Oracle are the operators that can be used to compute the decision score of the system. Kuncheva (2001) studied that among the six operators, system performance using average operator outperforms the performance using the other operators. Poh *et al.* (2002) validated this finding by modeling a score corrupted by noise as shown below.

Given sample i out of a total of N samples, let s_i be the observed measure score and $\hat{s}$ is a true measure score. η_i is denoted as noise. So that, the observed measure s_i can be written in term of $\hat{s}$ and η_i as in Eq. 4.

$$s_i = \hat{s} + \eta_i \tag{4}$$

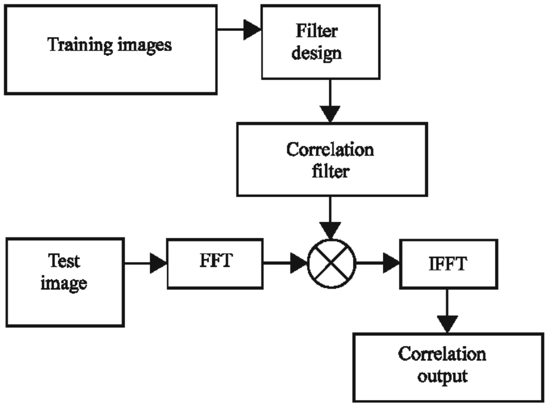

Fig. 1: Illustration of the correlation filter process

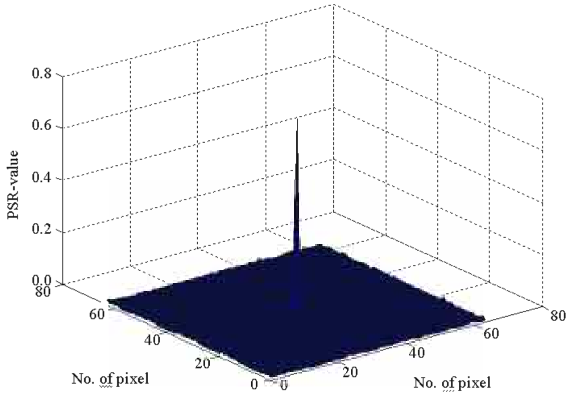

Fig. 2: Example of the correlation plane from the authentic class

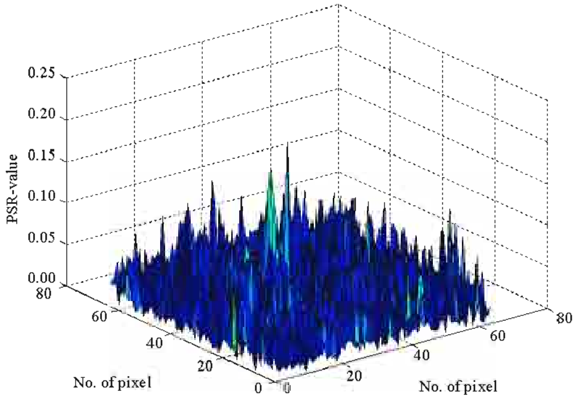

Fig. 3: Example of the correlation plane from the imposter class

The mean of the observed score, s_i, which denoted as $\bar{s}$ is then calculated as in Eq. 5:

$$\bar{s} = \frac{1}{N}\sum_{i=1}^{N} s_i \tag{5}$$

The expected value of s_i, denoted as $E\{\eta_i\}$ which is calculated as the mean of s_i, approximates the true value measure:

$$E\{s_i\} = \hat{s}. \tag{6}$$

Owing to the expected value of random noise η_i, $E(\eta_i)$ is always zero, the variance of the observed s is then calculated as in Eq. 7 by assuming the expected value of random noise η_i, $E\{\eta_i\}$ as zero.

$$\sigma_s^2 = \frac{1}{N}\sigma_\eta^2, \tag{7}$$

As a result, as shown in Eq. 7, the additive noise that imposed to the system can be diminished by a factor of the number of samples, N when N scores of a single modality biometric are averaged.

Feature extraction: The lipreading database used in this study is obtained from the Digit Database developed by Sanderson and Paliwal (2001). The database consists of video and corresponding audio reciting digits zero to nine. The video of each person is stored as a sequence of JPEG images with a resolution of 512×384 pixels. In order to locate the lip in a face, techniques for face detection and lip localization have been developed in this study.

In face detection task, 2 image processing steps are involved. The first step is to separate human skin regions from non-skin region using a color-based technique. Here, the chromatic colors in the absence of luminance are defined by normalizing the common RGB color representation as shown in Eq. 8 and 9 below.

$$r = \frac{R}{(R+G+B)} \tag{8}$$

$$b = \frac{B}{(R+G+B)} \tag{9}$$

Chromatic colors are effective to be implemented for segmenting skin color image and it can be represented by a Gaussian model, N(μ, C). A color distribution of skin color from different people clusters in small area of chromatic color space. The Gaussian model N(μ, C) with mean vector:

$$\mu = E[x] \tag{10}$$

and covariance matrix,

$$C = E[(x - \mu)(x - \mu)^T] \tag{11}$$

which are obtained from different skin color images has been used to determine the skin likelihood for any pixel of an image. The skin likelihood is computed as

$$P(r,b) = \exp[-0.5(x - m)^T C^{-1}(x - m)] \tag{12}$$

Where, $x = (r, b)^T$ is chromatic pair value of red and blue. The skin-likelihood image is transformed to the skin-segmented image (binary image) by setting a threshold value. The original image of the person, the skin-likelihood image and the skin segmented image are shown in Fig. 4.

The second step of face detection is the execution of the template matching technique. Here, we use a human face template to determine whether the skin region represents a face for the final decision. The cross-correlation value between the template face and the image that match to the skin region are then calculated. The final result of the face detection system is shown in Fig. 5.

For the lip localization task, hue and saturation color thresholding has been utilized to separate the lip region from the face skin. According to Matthews *et al.* (2002), the detection of the lip in hue and saturation color is much easier owing to its robustness under wide range of lip

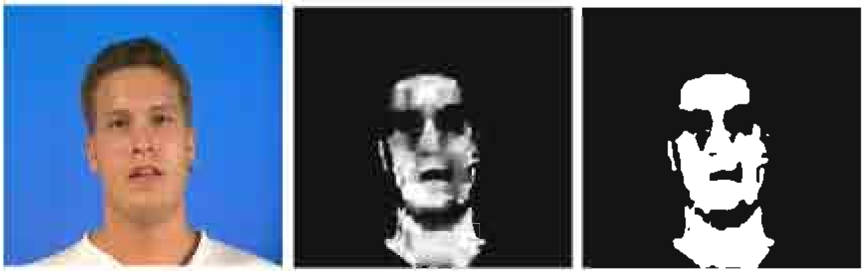

Fig. 4: Original image, skin-likelihood image and skin-segmented image

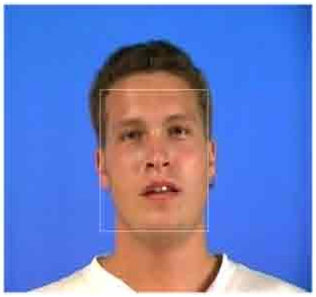

Fig. 5: Detected face

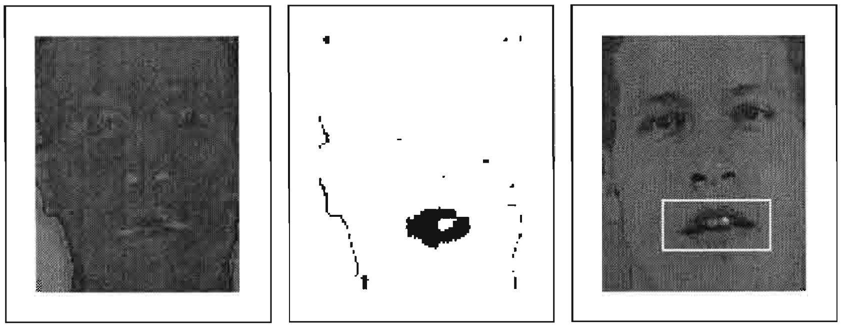

Fig. 6: Hue-saturation image (left), binary image (middle) and the localized lip region (right)

colors and varying illumination condition. From the hue-saturation image, a binary image is then computed by setting the threshold values, $H_0 = 0.04$ and $S_0 = 0.1$. By applying some morphological image processing such as opening, closing and noise removal, the largest blob is then determined as a lip region from the binary image. The lip region of 90×60 pixels is then extracted from the center of the largest blob. Figure 6 shows the hue-saturation image, the binary image obtained after applying the hue sand saturation color thresholding and the final output of this task.

The database and verification process: Our lipreading database consists of 20×25×41 = 20500 localized lip images. The first 20 frames of each sequence in the video from 25 people have been utilized; meanwhile, 41 sequences of frames for each person have been used for the purpose of this study. Training sequence which to be synthesized for the UMACE filter is chosen from one of the 41 sequences for each person consist of 20 training images. In our case, we have 25 filters which represent each person in the database. During the testing stage, we performed cross correlations of each person by using their corresponding filter with 800 authentic images (40 authentic sequences) and another 800×24 = 19200 imposter images (90 imposter sequences) from the other 24 persons. Figure 7 shows the details of the verification process in this experiment.

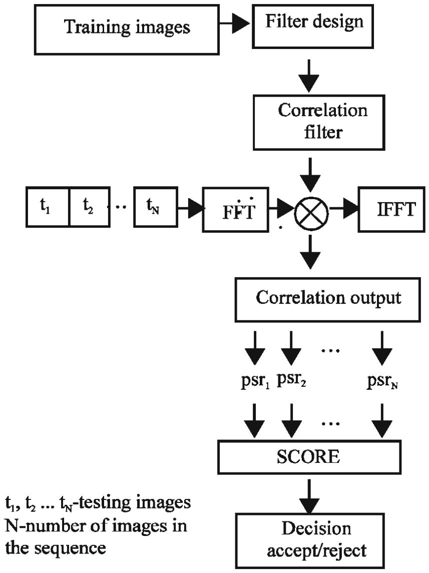

Fig. 7: Illustration of lipreading verification process

After verification process for each image in the sequence terminates, the SCORE value for the sequence is calculated so as to be used in the decision process.

SCORE value using multi sample fusion scheme via average operator is calculated as in Eq. 13.

$$\text{SCORE} = \sum_{i=1}^{n} \frac{\text{psr}_i}{n}\ ,\ \text{n - number of image} \qquad (13)$$

The verification decision is made by setting a SCORE threshold value, S_0 for each person. Then, the SCORE is compared to this threshold value, S_0 for the decision whether to accept or reject the person. If SCORE is greater than S_0, then the person is accepted as authentic person, otherwise as imposter person.

The False Acceptance Rate (FAR) and False Rejection Rate (FRR) are then calculated as defined below.

$$\text{FAR} = \frac{\text{No. of imposter person (SCORE} > S_0)}{\text{Total imposter person}} \qquad (14)$$

$$\text{FRR} = \frac{\text{No. of authentic person (SCORE} < S_0)}{\text{Total authentic person}} \qquad (15)$$

Then, overall performance is calculated by combining these two errors into Total Error Rate (TER) and Total Success Rate (TSR) as shown in Eq. 3 and 4.

$$\text{TER} = \left(\frac{\text{FAR + FRR}}{\text{Total No. of accesses}}\right)100\% \qquad (16)$$

$$\text{TSR} = 100\% - \left(\frac{\text{FAR + FRR}}{\text{Total No. of accesses}}\right)100\% \qquad (17)$$

RESULTS AND DISCUSSION

In the experiments, the performance of each person's UMACE filter was assessed by cross-correlating all the images for each sequence in the database and their corresponding PSR and SCORE values were computed and recorded. From the analysis, 20 out of 25 people in the database have been correctly verified (100% accuracy) and the error rate (TER) for those who are incorrectly verified is below than 1%.

Figure 8 shows SCORE performance of person II (TER equal to 0%). Solid line refers to the SCORE of the authentic sequence while dotted line to the imposter sequence. For person 9, SCORE values for all imposters are below 11 whereas for the authentic the values are above 20. Here, a wide margin of separation between the maximum SCORE value for imposters and the minimum SCORE value for the authentic offers a better potential for the verification process.

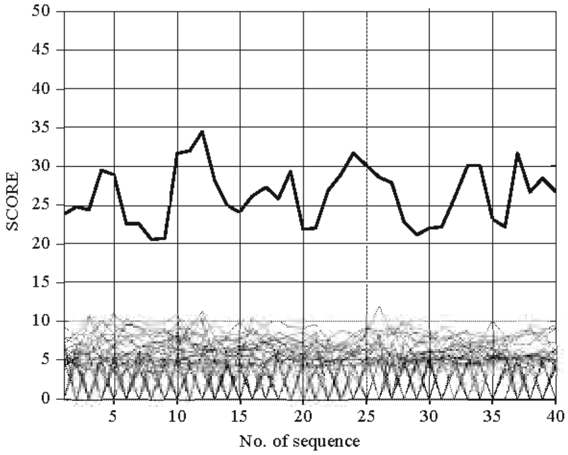

Fig. 8: The SCORE performance of person 11

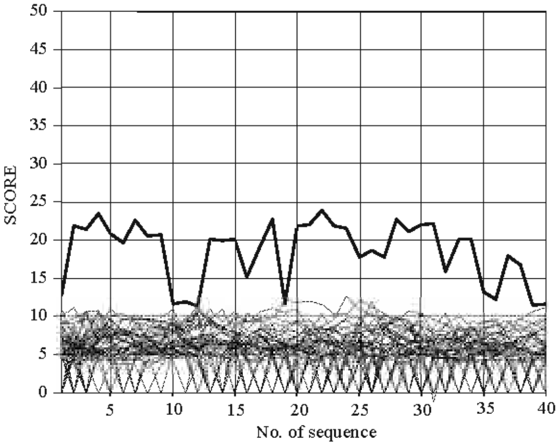

Fig. 9: The SCORE performance of person 18

We also illustrate the SCORE performance of person 18 who is incorrectly verified (TER equal to 0.8%) as shown in Fig. 9. A small gap between the maximum SCORE value for the imposters and the minimum SCORE value for the authentic is observed. The maximum SCORE value for the imposters sequence is 13 whereas for the authentic sequence, the minimum SCORE value is 11. It was scrutinized that, the images from sequence 10, 12, 19 and 39 have more variations due to the facial expression they made during speech production and some of the images in the sequence have been severely degraded by illumination. Owing to this problem, the training images were not sufficiently synthesized while designing the filter.

The overall performance (TSR) of single-sample system is evaluated as 92.29% from this experiment. The system performance increase to 7.31% after the implementation of multi-sample fusion approach yielding an overall performance of (TSR) 99.6%. Based on FRR and FAR percentages, the error percentages of 0.9% and 0.1%, respectively are observed from the multi-sample fusion system. From these encouraging results, we can conclude that the implementation of lipreading images as features and UMACE filters as classifier can be an alternative technique for speaker verification system. Our system offers simple features extraction process and UMACE filters are effective to perform well in the presence of the variations of image. The employment of multi-sample fusion technique to the system is also significant to enhance the reliably of the system performance. Further work will be assigned to the larger database and the development of multi-modal system under acoustically degraded condition.

REFERENCES

Broun, C.C., X. Zhang, R.M. Mersereau and M. Clements, 2002. Automatic speech reading with application to speaker verification. IEEE International Conference on Acoustics Speech and Signal Processing, 1: 685-688.

Campbell, J.P., 1997. Speaker recognition: A tutorial. Proc. IEEE, 85: 1437-1462.

Campbell, W.M., 2003. A SVM/HMM system for speaker recognition. Proceeding of IEEE on Acoustics Speech and Signal Processing, 2: 209-212.

Campbell, J.P., D.A. Reynolds and R.B. Dunn, 2003. Fusing high and low features for speaker recognition. Proc. EUROSPEECH, pp: 2665-2668.

Cheung, M.C., M.W. Mak and S.Y. Kung, 2004. Multi-sample data-dependent fusion of sorted score sequences for biometric verification. Proceeding of the IEEE Conference on Acoustics Speech and Signal Processing (ICASSP 04), pp: 229-232.

Faraj, M.I. and J. Bigun, 2006. Person verification by lip-motion. Proceeding of International Conference on Computer Vision and Pattern Recognition Workshop (CVPRW06), pp: 37-43.

Fox, N.A. and R.B. Reilly, 2004. Robust multi-modal person identification with tolerance of facial expression. Proceeding of IEEE International Conference on System, Man and Cybernetics, pp: 580-585.

Furui, S., 1997. Recent advances in speaker recognition. AVBPA97, pp: 237-251.

Hennecke, M.E., D.G. Stork and K.V. Prasad, 1996. Visionary Speech: Looking Ahead to Practical Speech reading Systems. Speech reading by Humans and Machines. Springer, pp: 331-349.

Kittler, J., G. Matas, K. Jonsson and M. Sanchez, 1997. Combining evidence in personal identity verification systems. Pattern Recog. Lett., 18 (9): 845-852.

Kittler, J., M. Hatef, R.P.W. Duin and J. Matas, 1998. On combining classifiers. Proceeding of the IEEE Trans. Pattern Anal. Mach. Intell., 20 (3): 226-239.

Kuncheva, L.I., 2001. A theoretical study on six classifier fusion strategies. Proceeding of the IEEE Transaction On Pattern Analysis and Machine Intelligence, pp: 348-353.

Matthews, I., J. Cootes, J. Bangham, S. Cox and R. Harvey, 2002. Extraction of visual features for Lipreading. IEEE Trans. Pattern Anal. Mach. Intell., 24 (2): 198-213.

Poh, N., S. Bengio and J. Korczak, 2002. A multi-sample multi-source model for biometric authentication. Proceeding. of the IEEE 12th Workshop on Neural Networks for Signal Processing, pp: 375-384.

Potamianos, G. and C. Neti, 2000. Improved ROI and within frame discriminant features for lipreading. Proceeding of the International Conference on Image Processing, pp: 250-253.

Potamianos, G., C. Neti, G. Gravier, A. Garg and A.W. Senior, 2003. Recent advances in the automatic recognition of audio-visual speech. Proc. IEEE, 91 (9): 1306-1326.

Reynolds, D.A., 2002. An overview of automatic speaker recognition technology. Proceeding of IEEE on Acoustics Speech and Signal Processing, 4: 4072-4075.

Samad, S.A., D.A. Ramli and A. Hussain, 2007a. Lower face verification centered on lip using correlation filters. Inform. Technol. J., 6 (8): 1146-1151.

Samad, S.A., D.A. Ramli and A. Hussain, 2007b. Person Identification Using Lip Motion Sequence. Lecture Notes in Computer Science (AI Series), Publisher Springer-Verlag Berlin Heidelberg, pp: 839-846.

Sanderson, C. and K.K. Paliwal, 2001. Noise compensation in a multi-modal verification system. Proceedings of International Conference on Acoustics, Speech and Signal Processing, pp: 157-160.

Savvides, M., B.V.K. Vijaya Kumar and P. Khosla, 2002. Face verification using correlation filters. Proceeding of 3rd IEEE Automatic Identification Advanced Technologies, pp: 56-61.

Savvides, M. and B.V.K. Vijaya Kumar, 2003. Efficient design of advanced correlation filters for robust distortion-tolerant face recognition. Proceeding of the IEEE Conference on Advanced Video and Signal Based Surveillance (AVSS'03), pp: 229-232.

Venkataramani, K. and B.V.K. Vijaya Kumar, 2003. Fingerprint Verification using Correlation Filters. System (AVBPA 2003), pp: 886-894.

Vijaya Kumar, B.V.K., 1992. Tutorial survey of composite filter designs for optical correlators. Applied Optics, 31: 4773-4801.

Wark, T., D. Thambiratnam and S. Sridharan, 1997. Person authentication using lip information. Proceeding of IEEE on Speech and Image Technologies for Computing and Telecommunications (TENCON), pp: 153-156.

Wark, T. and S. Sridharan, 1998. A syntactic approach to automatic lip feature extraction for speaker identification. IEEE International Conference on Acoustics Speech and Signal Processing, 6: 3693-3696.

Wark, T., S. Sridharan and V. Chandran, 1999. The use of speech and lip modalities for robust speaker verification under adverse conditions. IEEE International Conference on Acoustics Speech and Signal Processing, 6: 3061-3064.