An Adaptive Fusion using SVM based Audio Reliability Estimation for Biometric Systems

Dzati Athiar Ramli, *Member, IAENG*   Salina Abdul Samad   Aini Hussain

***Abstract*—Performances of speaker verification systems are superb in clean noise-free conditions but the reliability of the systems drop severely in noisy environments. Fusion of audio and visual information is one of the solutions to this limitation. However, this approach requires appropriate weighting for each biometric trait when the systems are implemented under inconsistent conditions. In this study, we propose a novel approach by introducing Support Vector Machine (SVM) as indicator system for audio reliability estimation. This approach directly validate the quality of the incoming (claimant) speech signal so as to adaptively change the weighting factor for fusion of both subsystems scores. It is important to priory check the speech signal quality because unreliable speech data give incorrect scores hence affect the accuracy of the total scores of the fusion systems. The effectiveness of this approach has been experimented to a multibiometric verification system that employs lipreading images as visual features. This verification system uses SVM as a classifier for both subsystems. Principle Component Analysis (PCA) technique is executed for visual features extraction while for the audio feature extraction; Linear Predictive Coding (LPC) technique has been utilized. In this study, we found that the SVM indicator system is able to determine the quality of the speech signal up to 99.66%. We then observed that by using the proposed adaptive fusion system, EER percentage in noisy condition (10dB) has been decreased to 0.27% compared to 9.3% for non-adaptive fusion system and 51.13% for audio only system.**

***Index Terms*—audio visual system, biometric verification system, reliability estimation, Support Vector Machine.**

I. Introduction

Biometric speaker verification is a technology that utilizes behavioral and physiological information of speech signal for the purpose of authentication of individual for identity claim.

Manuscript received February 28, 2009. This work was supported by Fundamental Research Grant Scheme, Malaysian Ministry of Higher Education, FRGS UKM-KK-02-FRGS0036-2006 and Science Fund, Malaysian Ministry of Sciences, Technology and Innovation 01-01-02-SF0374.

Dzati Athiar Ramli is with School of Electrical & Electronic Engineering, Universiti Sains Malaysia, MALAYSIA and currently as PHD candidate at Department of Electrical, Electronic & System Engineering, Engineering Faculty, Universiti Kebangsaan Malaysia, MALAYSIA, e-mail: dzati@vlsi.eng.ukm.my, dzati@eng.usm.my).

Salina Abdul Samad is with the Department of Electrical, Electronic & System Engineering, Engineering Faculty, Universiti Kebangsaan Malaysia, MALAYSIA (e-mail: salina@vlsi.eng.ukm.my).

Aini Hussain is with the Department of Electrical, Electronic & System Engineering, Engineering Faculty, Universiti Kebangsaan Malaysia, MALAYSIA (e-mail: aini@vlsi.eng.ukm.my).

According to [1], [2], the advantages of using speech signal trait for biometric systems are that the signal is natural and easy to produce, requiring little custom hardware, has low computation requirement and is highly accurate (in clean noise-free conditions). However, in uncontrolled conditions, the reliability of the system drops severely as the signal to noise ratio (SNR) of the speech signal decreases. This becomes the main problem when utilizing speech signals for biometric systems. Furthermore, since voice is categorized as a behavioral signal, the signal is likely to vary in time due to the change of speaking rates, health and emotional conditions of speakers. Different microphones and channels also affect the accuracy of the system performance. Consequently, the implementation of biometric systems has to appropriately discriminate the biometric features from one individual to another, and at the same time, the systems also need to deal with the distortions of the features.

One of the solutions to overcome these limitations is by implementing fusion approach to the biometric system [3]. Reference [4] reported the fusion of scores produced independently by speaker recognition system and face recognition system using a weighted merged score. The optimal weight was found by maximizing the performance of the integrated system on one of the available training sets. The identification of 51% was achieved for the speech only system and 92% for the face only system. Performance of the integration system using the optimal weight is observed up to 95%.

In another case, a weighted product approach to fuse two voice features i.e. static and dynamic and three face features i.e. eye, noise and mouth was evaluated [5]. The tan-estimators were used for score normalization and weighted geometric average was used for score combination. The correct identification rate of the integrated system is 98% which represents a significant improvement with respect to the 88% and 91% rates provided by the speaker and face recognition systems respectively. Reference [5] combined different biometric cues i.e. voice, lip motion and face image. The EER performance of face recognition, voice recognition and integrated face and voice recognition are obtained as 3%, 3.4% and 1.5% from this experiment.

Reference [6] integrated the scores of speech and lip modality using weighted summation fusion. The performance of the integrated system outperformed each subsystem and reduced the false acceptance rate of the speech subsystem from 2.3% to 0.5%. In another experiment,

ISBN: 978-988-17012-5-1

information from speaker verification system (SVS) and profile verification system (PVS) using a weighted summation fusion was combined [7]. A weighting factor w is fixed between 0 and 1. Fusion system using $w = 0.33$ gives verification rates of 95.57% (40dB) and 72.82% (5dB) while performances of 90.31% (40dB) and 78.75% (5dB) are then observed using $w = 0.5$.

In [9], a novel fuse-HMM that integrates the audio and visual features of speech was reported. In this method, the learning algorithm maximizes the two HMMs separately and consequently fuse the HMM by Bayesian fusion method. The experimental results showed that the fuse-HMMs constantly performed better than the unimodal method under clean and low noise conditions. But under stronger noise level, the performance of the fusion systems is worse compared to the speech only system.

Multistage information fusion by taking both feature fusion and decision fusion approach was implemented in [10]. The study observed that the multistage system achieves significant improvement over both feature fusion and decision fusion system at different SNR levels.

Studies on audio reliability estimation are also reported in literatures. This method is performed either relying on the statistics-based reliability measure or directly based on the quality of the speech signal. Here, the weight for fusion scheme is adapted correspondingly to the quality of the current input (claimant) speech signal instead of using the optimum weight that is estimated from the available training set. This approach is more advantageous especially when the system is implemented in uncertain environment conditions.

Two methods have been proposed for the statistics based reliability measures i.e. entropy of a posteriori probabilities and dispersion of a posteriori probabilities. The reliability information can be obtained by the shape of a posteriori probabilities distribution of HMM states, GMM and MLP as studied in [11], [12] and [13], respectively. A high entropy interprets low confidence hence signifies very unreliable input. Consequently, a mapping function between the entropies and the corresponding weight is calculated.

On the other hand, study on reliability estimation based on the quality of the speech signal was reported in [13]. This study described the use of voicing index as audio reliability measure. Implementation of the degree of voicing index as reliability measure is also reported in [14].

In this study, we propose a novel approach by introducing Support Vector Machine as indicator system for audio reliability measure. The development of this system is made up of 3 modules i.e. an audio front-end module, a visual front-end module and a fusion and verification module. For audio front-end module, a vector of LPC coefficients is computed from the autocorrelation vector using Durbin recursion method. The LPC-derived cepstral coefficients (cepstrum) are then extracted.

For the visual front-end module, lipreading features are employed to the system. Lipreading features are the sequence of lip images while the speaker utters the words for example, zero to nine. The advantages of utilizing lipreading features together with speech signals include the simple process of data collection and the cost effective factor since they can be simultaneously captured using the same hardware, i.e., digital video camera. In addition, the use of lip features, compared with face, can also minimize the storage capacity and increase the speed of computation as well. Several researches using lip information as features to recognition systems have been reported. As in [15], shape and intensity information from a person's lip were used in a speaker recognition system. The utilization of geometric dimension such as height, width and angle of speaker's mouth as features was also investigated [16]. Apart from lip contour-based features, pixel-based features i.e. Discrete Cosine Transform (DCT) has also been experimented as features for person recognition in [17].

Finally, two tasks are executed for the fusion and verification module. In the first task, the SVM indicator system is developed for audio reliability measure. The weighting factor for score integration is then decided in the second task accordingly to the audio reliability estimation result. The overall architecture of the proposed adaptive weight fusion system is illustrated in Fig. 1.

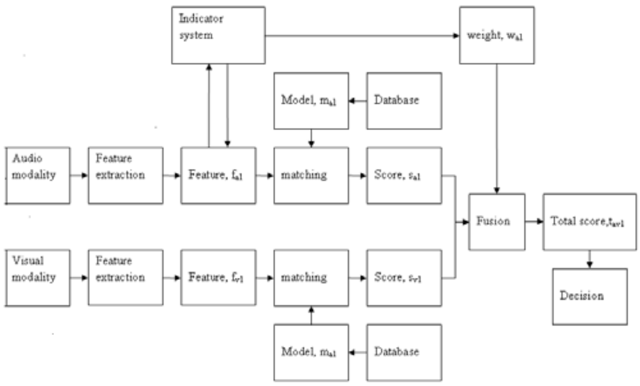

Fig. 1 Adaptive weight fusion systems

The first objective of this study is to examine the performance of the SVM indicator system for audio reliability estimation. Secondly, it evaluates the performance of the proposed adaptive fusion scheme in clean noise-free condition and under noisy conditions. Different levels of signal to noise ratios (SNRs) of speech signals, ranging from clean to 10dB are experimented so as to simulate the real life conditions. Finally, we also compare the performances of the adaptive fusion system with non-adaptive fusion system and audio only system at different levels of SNR.

The database used in this study is the Audio-Visual Digit Database (2001) [18]. The database consists of video and the corresponding audio recording of people reciting digits zero to nine. The video recording of each person is stored as a sequence of JPEG images with a resolution of 512 x 384 pixels while the corresponding audio recording provided is a monophonic, 16 bit, 32 kHz, WAV format. This paper is organized as follows. In section II, the Support Vector Machine classifier is explained. Section III and IV describe the audio verification module and visual verification modules, respectively. Fusion and verification module is then represented in section V. In section VI, we discuss our results and finally, some conclusions are summarized in section VII.

II. SUPPORT VECTOR MACHINE CLASSIFIER

Support vector machine (SVM) classifier in its simplest form, linear and separable case is the optimal hyper plane that maximizes the distance of the separating hyper plane from the closest training data point called the support vectors [19], [20].

From [19], the solution of a linearly separable case is given as follows. Consider a problem of separating the set of training vectors belonging to two separate classes,

$$D = \left\{ \left(x^1, y^1\right), \ldots \left(x^L, y^L\right) \right\}, \qquad x \in \Re^n, y \in \{-1, -1\} \tag{1}$$

with a hyperplane,

$$\langle w, x \rangle + b = 0 \tag{2}$$

The hyperplane that optimally separates the data is the one that minimizes

$$\phi(w) = \frac{1}{2}\|w\|^2 \tag{3}$$

which is equivalent to minimizing an upper bound on VC dimension. The solution to the optimization problem (3) is given by the saddle point of the Lagrange functional (Lagrangian)

$$\phi(w, b, \alpha) = \frac{1}{2}\|w\|^2 - \sum_{i=1}^{L} \alpha_i \left(y^i \left[\left\langle w, x^i \right\rangle + b \right] - 1 \right) \tag{4}$$

where α are the Lagrange multipliers. The Lagrangian has to be minimized with respect to w, b and maximized with respect to $\alpha \geq 0$. Equation (3) is then transformed to its dual problem. Hence, the solution of the linearly separable case is given by,

$$\alpha^* = \arg\min_{\alpha} \frac{1}{2} \sum_{i=1}^{L} \sum_{j=1}^{L} \alpha_i \alpha_j y_i y_j \left\langle x_i, x_j \right\rangle - \sum_{k=1}^{L} \alpha_k \tag{5}$$

with constrains,

$$\alpha_i \geq 0, \quad i = 1, \ldots, L \quad \text{and} \quad \sum_{j=1}^{L} \alpha_j y_j = 0 \tag{6}$$

Subsequently, consider a SVM as a non-linear and non-separable case. Non-separable case is considered by adding an upper bound to the Lagrange multipliers and non-linear case is considered by replacing the inner product by a kernel function. The solution of the non-linear and non-separable case is given as:

$$\alpha^* = \arg\min_{\alpha} \frac{1}{2} \sum_{i=1}^{L} \sum_{j=1}^{L} \alpha_i \alpha_j y_i y_j K\left(x_i, x_j\right) - \sum_{k=1}^{L} \alpha_k \tag{7}$$

with constrains,

$$0 \leq \alpha_i \leq C, \quad i = 1, \ldots, L \tag{8}$$

$$\text{and} \quad \sum_{j=1}^{L} \alpha_j y_j = 0 \;\; x(t) = (s(t) - 0.95) * x(t-1) \tag{9}$$

Non-linear mappings (kernel functions) that can be employed are polynomials, radial basis functions and certain sigmoid functions.

III. AUDIO FRONT-END SUBSYSTEM

Linear Predictive Coding is a time domain analysis that approximates a speech sample as a linear combination of past speech samples. A unique set of predictor coefficients are determined by minimizing the sum of the squared differences between the actual speech samples and the linearly predicted ones [21], [22]. The steps for obtaining Linear Predictive Coding (LPC) are summarized in Fig. 2. The parameter values that have been used at each stage of the experiment are also indicated. A set of feature vector computed from each frame consists of 14 cepstrum coefficients.

Experiments for clean condition systems use clean data for training and testing. For the noisy condition experiment, clean data are used for training while for testing; the speech signal data have been corrupted by decreasing the signal to noise ratio (SNR) into 30dB, 20dB and 10dB using white Gaussian noise.

For both clean and noisy condition experiments, we fix the number of client training data to 20. For the classifier to be trained discriminatively, each speaker is trained on 20 client data and 720 (20x36) imposter data. Consequently, each speaker is tested on 40 client test data and 1440 (40x36) imposter test data. Since we have 37 speakers in the database, we have constructed 37 SVM models for the experiment.

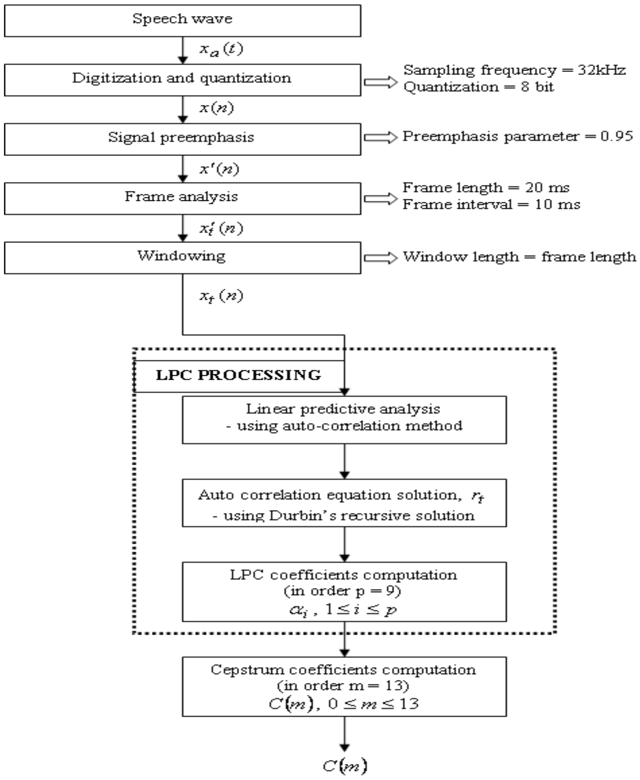

Fig. 2 LPC processing

IV. VISUAL FRONT-END SUBSYSTEM

In order to locate the lips on a face, techniques for face detection and lip localization have been used in this study [23], [24]. In the first task, we implement a color-based technique and template matching algorithm to segment human skin regions from non-skin color. For the lip localization task, hue/saturation color thresholding has been employed in order to differentiate the lip area from the face [23], [24]. As demonstrated in [25], the detection of the lip in hue/saturation color is much easier owing to its robustness under a wide range of lip colours and varying illumination condition. From the hue-saturation image, a binary image is then computed followed by morphological image processing so as to determine lip region. The lip regions of 64 x 64 pixels are then extracted for evaluation. Our lipreading database has 22200 images in total size 64x64 pixels from 37 persons. For each person, 60 sequences of images (with 10 images per sequence) have been utilized.

Consequently, Principle component analysis (PCA) technique or also known as Karhunen-Loeve method is used for dimensionality reduction. This statistical method aims to obtain an optimum linear subspace from a covariance matrix of a set of samples [26]. This technique executes linear projection on the original samples that maximizes the scatter (variance) of all projected samples. This technique is beneficial for reducing storage capacity because the projected features are presented in a lower dimensionality space compared to the original sample space. The overall feature extraction process using PCA technique is summarized in Fig. 3.

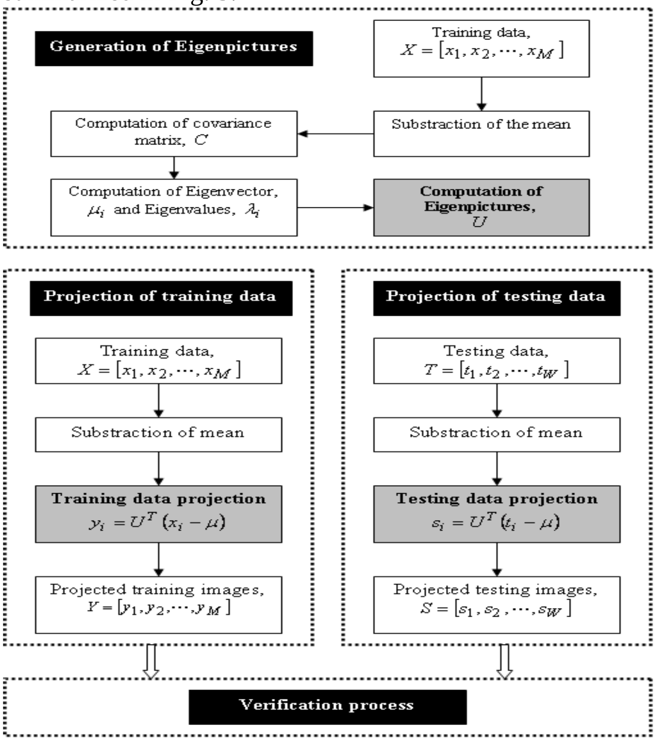

Fig. 3 Feature extraction using PCA processing

Theory of PCA technique for feature extraction can be simply stated as follows. Given a set of N sample images x_i, $i = 1,2,\cdots,M$ where each image in the set is lethographically re-ordered in L^2 dimensional space and belongs to one of the c classes $\{C_1, C_2, \cdots, C_c\}$. By considering a linear transformation mapping, the original sample in L^2 dimensional space are then transformed into a P-dimensional feature space, where $P << M << L^2$. The new transformed features y_i, $i = 1,2,\cdots,M$ is known as subspace and the process of transforming is called projection. In PCA, the transformation process is executed by the following linear transformation:

$$y_i = U^T x_i, \quad i = 1,2,\cdots,M \tag{10}$$

where $U \in \Re^{L^2 \text{ x } P}$ represents matrix of Eigen pictures in $L^2 \text{x} P$ and P corresponding to the P largest Eigen values.

The transformed lip features are then used for the verification process using SVM as classifier. In this experiment, we vary the number of training data from 3, 6, 10 and 20 so as to evaluate the performance of the systems using a different number of training data. For the classifier to be trained discriminatively, each speaker is trained on 3, 6, 10 or 20 client data and 720 (20x36) imposter data. Consequently, each speaker is tested on 40 client test features and 1440 (40x36) imposter test features. As we have 37 speakers in the database, we have constructed 37 SVM models for the experiment.

V. Fusion and verification system

A. *SVM indicator system*

The architecture of the score level fusion using adaptive weight fusion scheme is illustrated in Fig. 4. Speech quality measurement is done by developing an indicator system which is based on SVM classification technique. By modeling the clean data features as sample type +1 and the noisy data features as sample type -1, the system is able to discriminate the incoming speech signal either as high quality or low quality speech signal. Modeling data are taken from the training data set (enrollment). After the speech quality measurement process is completed, the system will decide the weight for the fusion process.

The indicator system is constructed to differentiate clean speech signal (high quality) from 30dB, 20dB and 10dB SNR speech signal (low quality). We have used 2960 training data and 5920 testing data for this task. This system is capable to achieve 99.66% accuracy.

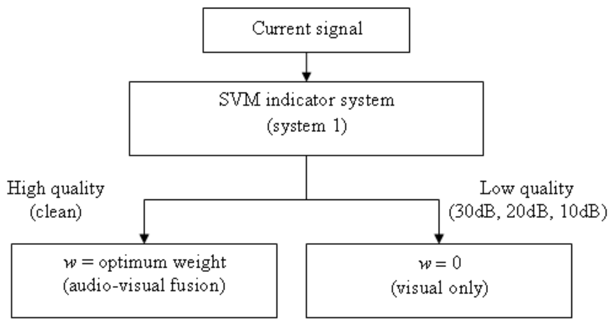

Fig. 4 The architecture of the score level fusion using adaptive weight fusion scheme

B. *Fusion system Verification*

The fusion system is a soft fusion system that uses raw scores from audio and visual subsystems. The first task is to normalize the values from each of the subsystems using min-max normalization by placing them in the [0,1] interval. The normalized values are then combined by using a weighted summation fusion as shown by the equation below.

$$F = (1 - w)V_{lip} + wA_{speech} \tag{11}$$

where V_{lip} is the score from visual subsystem, A_{speech} is the score from audio subsystem, F is total weighted fusion score and W is a weighting factor which varied between 0 and 1.

Before the fusion process takes place, each audio testing data (current speech signal) is first checked for its quality by the indicator system and the weight for the particular data is then determined. By using the weight defined from the indicator system, the normalized scores from both subsystems are then fused using (11). In order to calculate the optimum weight, w_{opt}, w is varied from 0 to 1 in steps of 0.2. The overall performance in each step is then evaluated and the optimum weight, w_{opt} is defined at which the weight, w give the highest performance. We observed that the error curve performances hit the lowest point at weight equal to 0.4.

VI. Results and Discussions

We first discuss the performance of the visual only systems that have been experimented based on 20, 10, 6 and 3 training data. By increasing the number of training data to the systems, a great improvement in GAR is observed. The 100% GAR for 3, 6, 10 and 20 training data systems is found when the FAR is equal to 35%, 4%, 3% and 0.2%, respectively. System performances based on EER are shown in Table I.

Table I EER performances for visual only systems

No of training data	20	10	6	3
EER	0.27	0.94	1.15	2.7

Performances of the adaptive fusion systems have been experimented in clean, 30db SNR, 20dB SNR and 10dB SNR conditions. Performances of non-adaptive fusion systems and audio only systems have also experimented in this study. For the non-adaptive fusion systems, the optimum weight $w = 0.4$ has been used for fusion of the scores.

Fig. 5 illustrates the performances of the verification systems using adaptive weight fusion and non-adaptive weight fusion in clean condition.

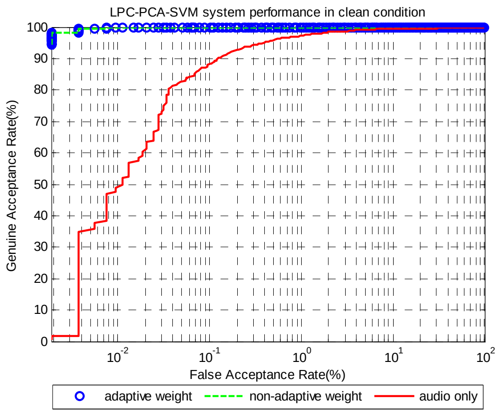

Fig. 5 System performance in clean condition

The performance of the audio only system in clean condition is also illustrated for comparison. The 100% GAR performance is evaluated at FAR is equal to 0.004% for the adaptive weight fusion system and non-adaptive weight fusion system compared to 35% GAR performance for audio only system at the same percentage of FAR. It is observed that the audio only system reaches 100% GAR at FAR of 4%. System performances based on EER are observed as 0.067%, 0.067% and 1.79% for adaptive weight fusion system, non-adaptive weight fusion system and audio only system, respectively.

Performances of the verification systems using adaptive weight fusion and non-adaptive weight fusion based on 30dB SNR data are given in Fig. 6. The performance of the audio only system based on 30dB data is also illustrated for comparison. 100% GAR performance is evaluated for the adaptive weight fusion system at FAR equal to 0.3% compared to FAR equal to 1% for non-adaptive weight fusion system. The GAR performance at the FAR equal to 0.3% for non-adaptive weight fusion is 96% meanwhile GAR performance for audio only system is 18% at the same FAR percentage. System performances based on EER are observed as 0.29%, 0.87% and 17.02% for adaptive weight fusion system, non-adaptive weight fusion system and audio only system, respectively.

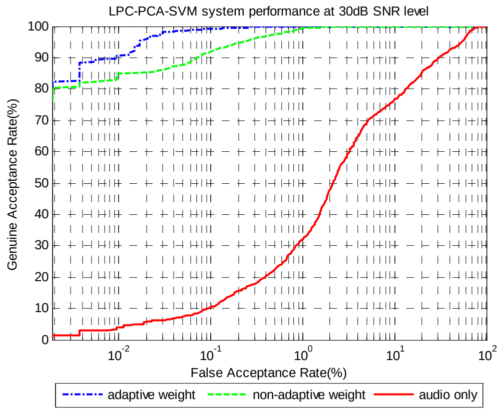

Fig. 6 System performances at 30dB SNR level

Fig. 7 illustrates the performances of the verification systems using adaptive weight fusion and non-adaptive weight fusion based on 20dB SNR data.

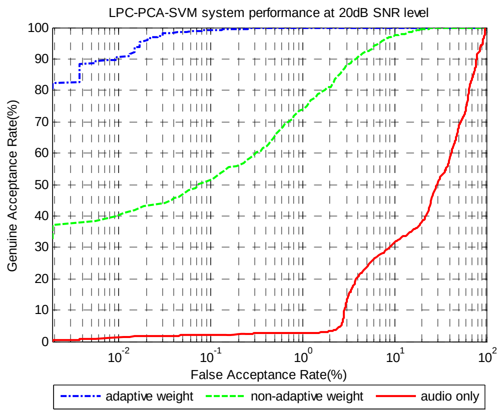

Fig. 7 System performances at 20dB SNR level

The performance of the audio only system based on 20dB data is also illustrated for comparison. The 100% GAR is evaluated at FAR is equal to 0.3% for the adaptive weight fusion system compared to 60% GAR for non-adaptive weight fusion system at the same FAR percentage. In contrast, the non-adaptive weight fusion system reaches 100% GAR at FAR equal to 11%. At the FAR 0.3%, the audio only system attains 4% GAR. System performances based on EER are observed as 0.27%, 6% and 40.75% for adaptive weight fusion system, non-adaptive weight fusion system and audio only system, respectively.

Performances of the verification systems using adaptive weight fusion and non-adaptive weight fusion based on 10dB SNR data are given in Fig. 8. Performance of audio only system based on 10dB data is also given for evaluation. The 100% GAR performance for adaptive weight fusion system is evaluated at FAR equal to 0.3% while for non-adaptive weight fusion system is found at FAR of 35%. In contrast, the performance at FAR of 0.3% for non-adaptive weight fusion is 48% GAR meanwhile GAR performance for audio only system is evaluated as 0% at the same FAR percentage. System performances based on EER are observed as 0.27%, 9.3% and 51.14% for adaptive weight fusion system, non-adaptive weight fusion system and audio only system, respectively.

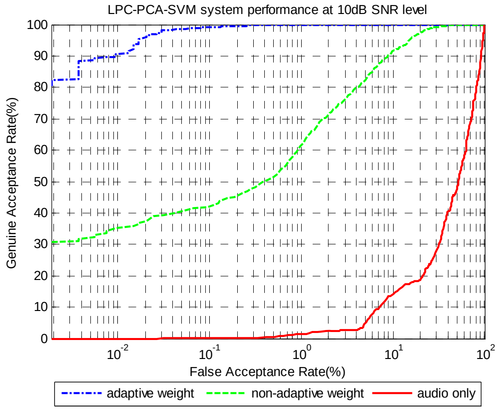

Fig. 8 System performances at 10dB SNR level

VII. Conclusions

The performances of the adaptive weight fusion systems, non-adaptive weight fusion systems and audio only systems at different SNR levels have been reported for comparison in this paper. The advantage of using the adaptive weight fusion approach instead of employing non-adaptive weight fusion is to avoid unreliable scores to be fused together in fusion systems that can spoil the accuracy of the total scores. We conclude that the adaptive weight fusion systems always outperform the other systems. By using the adaptive weight fusion approach, the performances of the verification systems can be further enhanced when high quality speech signal is obtained. Besides, in corrupted speech signal environment, the system performances can still be maintained by adjusting the fusion weight by using the visual only systems. However, the effectiveness of this approach depends on the performance of the indicator system as audio reliability estimation and visual only verification system. Future work will be devoted on all SNR levels and different types of noises.

References

[1] J.P. Campbell, “Speaker Recognition: A Tutorial,” Proceeding of the IEEE 85, 1997, pp.1437-1462.
[2] D.A. Reynolds, “An overview of Automatic Speaker Recognition Technology,” Proceeding of IEEE on Acoustics Speech and Signal Processing 4, 2002, pp. 4072-4075.
[3] D.A. Ramli, S.A. Samad, S.A and A. Hussain, “Score Information Decision Fusion using Support Vector Machine for a Correlation Filter Based Speaker Authentication System,” Lecture Notes in Computer Science (Artificial Intelligent Series), Publisher Springer-Verlag Berlin Heidelberg, 2008, Vol 53, pp.235-242.
[4] R. Brunelli, D. Falavigna, L. Stringa and T. Poggio, “Automatic Person Recognition by Using Acoustic and Geometric,” Machine Vision & Applications 8, 1995, pp. 317-325.
[5] R. Brunelli and D. Falavigna, “Personal Identification using Multiple Cue,” IEEE Trans. on Pattern Analysis and Machine Intelligence, vol. 17, no. 3, 1995, pp. 955–966.
[6] U. Dieckmann, P. Plankensteiner and T. Wagner, “SESAM: A Biometric Person Identification System using Sensor,” Pattern Recognition Letters, 1997, pp. 827–833.
[7] P. Jourlin, J. Luettin, D. Genoud and H. Wassner, “Integrating Acoustic and Labial Information for Speaker Identification and Verification,” Proceeeding 5th European Conference Speech Communication and Technology, vol. 3, 1997, pp. 1603–1606.
[8] C. Sanderson and K.K. Paliwal, “Multi-Modal Person Verification System Based on Face Profile and Speech,”Fifth International Symposium on Signal Processing and its Applications, 1999, pp. 947-950.
[9] H. Pan, Z.P. Liang and T.S. Huang, “Fusing Audio and Visual Features of Speech,” Proceeding of IEEE International Conference on Image Processing, 2000, vol 3, pp. 214-217.
[10] S.M. Chu, V.L.E. Marcheret, C. Neti and G. Potamianos, “Multistage Information Fusion for Audio-Visual Speech Recognition,” Proceeding of IEEE International Conference on Multimedia and Expo, 2004, pp. 1651-1654.
[11] M. Gurban and J.P. Thiran, “Using entropy as a stream reliability estimate for audio-visual speech,” 16th European Signal Processing Conference, 2008, to be published.
[12] G. Potamianos and C. Neti, “Stream Confidence Estimation for Audio-Visual Speech,” Procceding of the International Conference on Spoken Language, vol III, 2000, pp. 746-749.
[13] M. Heckmann, F. Berthommier and K. Kroschel, “Noise Adaptive Stream Weighting in Audio-Visual Speech, EURASIP Journal on Applied Signal Processing, 2002, pp. 1260-1273.
[14] G. Chetty and M. Wagner, “Robust Face-Voice Based Speaker Verification using Multilevel,” Image and Vision Computing, vol. 26, issue 9, 2008, pp. 1249-1260.
[15] T. Wark and S. Sridharan, “A Syntactic Approach to Automatic Lip Feature Extraction for Speaker Identification,” IEEE International Conference on Acoustics Speech and Signal Processing 6, 1998, pp. 3693-3696.
[16] C.C. Broun, X. Zhang, R.M. Mersereau and M. Clements, “Automatic Speechreading with Application to Speaker Verification,” IEEE International Conference on Acoustics Speech and Signal Processing1, 2002, pp. 685-688.
[17] N.A. Fox and R.B. Reilly, “Robust Multi-Modal Person Identification with tolerance of Facial expression,” Proceeding of IEEE International Conference on System, Man and Cybernetics, 2004, pp.580-585.
[18] C. Sanderson and K.K. Paliwal, “Noise Compensation in a Multi-Modal Verification System,” Proceeding of International Conference on Acoustics, Speech and Signal Processing, 2001, pp. 157-160.
[19] S.R. Gunn, “Support Vector Machine for Classification and Regression,” Technical Report, University of Southampton, 2005.
[20] V. Wan and W.M. Campbell, “Support Vector Machines for Speaker Verification and Identification,” Proceeding of Neural Networks for Signal Processing, 2000, pp. 775-784.
[21] L. R. Rabiner and B.H. Juang, “Fundamental of Speech Recognition Liveness Verification in Audio-Video Speaker Authentication,” in *Prentice-Hall International*, United State: Prentice-Hall International, 1993.
[22] S. Furui, “Cepstral Analysis Technique for Automatic Speaker Verification,” IEEE Transaction Acoustic Speech Signal Processing, vol. 29, no. 2, 1981, pp. 254-272.
[23] G. Chetty and M. Wagner, “Liveness Verification in Audio-Video Speaker Authentication,” Proceeding of International Conference on Spoken Language Processing ICSLP 04, 2004, pp. 2509-2512.
[24] G. Chetty and M. Wagner, “Automated Lip Feature Extraction for Liveness Verification in Audio-Video Authentication,” Proceeding of Image and Vision Computing, 2004, pp. 17-22.
[25] I. Matthews, J. Cootes, J. Bangham, S. Cox and R. Harvey, “Extraction of Visual Features for Lipreading,” IEEE Trans. on Pattern Analysis and Machine Intelligence, vol. 24, no. 2, 2002, pp. 198–213.
[26] M. Kirby and L. Sirovich, “Application of the Karhunen-Loeve Procedure for the Characterisation of human,” IEEE Transaction on Pattern Analysis and Machine, vol. 12, no. 1, 1990, pp. 103-108.

ISBN: 978-988-17012-5-1