Chapter 15
An Adaptive Multibiometric System for Uncertain Audio Condition

Dzati Athiar Ramli, Salina Abdul Samad, and 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. 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 subsystem scores. 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%. At 10 dB SNR, EER performances are observed as 51.13%, 9.3%, and 0.27% for audio only system, fixed weighting system and adaptive weighting system, respectively.

Keywords Biometric verification system · reliability estimation · support vector machine

15.1 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. 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

D.A. Ramli (✉), S.A. Samad, and A. Hussain
Department of Electrical, Electronic & System Engineering, Engineering Faculty,
Universiti Kebangsaan Malaysia, 43600, Bangi, Selangor, Malaysia
e-mail: dzati@vlsi.eng.ukm.my; salina@vlsi.eng.ukm.my; aini@vlsi.eng.ukm.my

S.-I. Ao and L. Gelman (eds.), *Electronic Engineering and Computing Technology*,
Lecture Notes in Electrical Engineering 60, DOI 10.1007/978-90-481-8776-8_15,

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. One of the solutions to overcome these limitations is by implementing fusion approach to the biometric system [3].

Research on fixed weighting fusion approach can be found in [4]. This study 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. 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. Reference [5] combined different biometric cues i.e. voice, lip motion and face image. Reference [6, 7] integrated the scores of speech and lip modality using weighted summation fusion. In another experiment, information from speaker verification system (SVS) and profile verification system (PVS) using a weighted summation fusion was combined [7]. In [9], fuse-HMM that integrates the audio and visual features of speech were 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.

For the adaptive weighting fusion approach, the reliability estimation of the current speech signal is performed either relying on the statistic based measure or directly based on the quality of the speech signal. 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 [11–13], 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 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 three modules i.e. an audio front-end module, a visual front-end module and a fusion module. For audio front-end module, a vector of LPC coefficients is computed from the autocorrelation vector using Durbin recursion method. 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. 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]. The overall architecture of the proposed adaptive weighting fusion system is illustrated in Fig. 15.1.

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 × 384 pixels while the corresponding audio recording provided is a monophonic, 16 bit, 32 kHz, WAV format. For the purpose of evaluating the systems in noisy conditions, the clean testing audio data are corrupted into 30, 20 and 10 dB SNR data by using the simulated additive white Gaussian noise (AWGN). Due to the objective of this research is to investigate the biometric system in uncertain audio condition, no artificial degradation was imposed to the visual data. However, some natural challenges such as facial expression, pose and illumination invariant are occurred within the sequence of the images and from session to session.

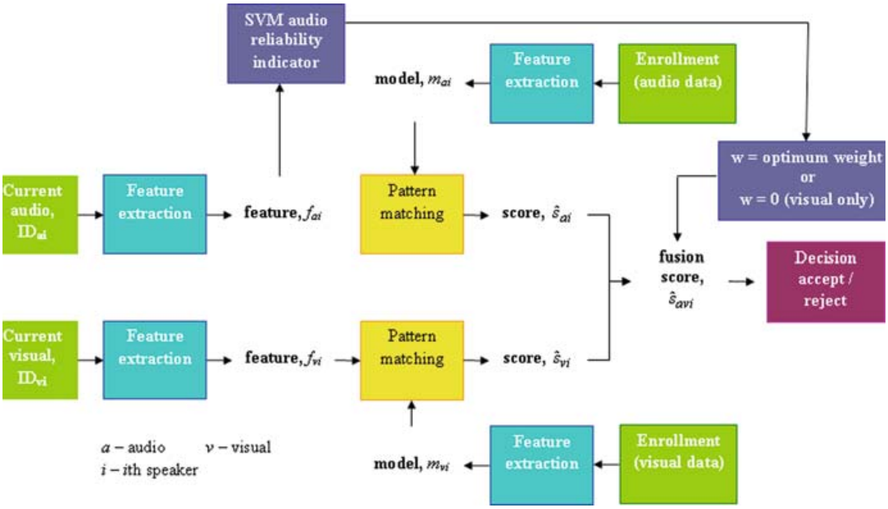

Fig. 15.1 An adaptive multibiometric system

15.2 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,

$$D = \left\{ \left(x^1, y^1\right), \ldots \left(x^L, y^L\right) \right\}, \ldots x \in \Re^n, y \in \{-1, -1\} \qquad (15.1)$$

with a hyperplane, $\langle w, x \rangle + b = 0$. The hyperplane that optimally separates the data is the one that minimizes

$$\phi(w) = \frac{1}{2} \|w\|^2 \qquad (15.2)$$

which is equivalent to minimizing an upper bound on VC dimension. The solution to the optimization problem, Eq. (15.2) is given by the saddle point of the Lagrange functional (Lagrangian)

$$\phi(w, b, \alpha) = \frac{1}{2} \|w\|^2 - \sum_{i=1}^{L} \alpha_i \left(y^i \left[\left\langle w, x^i \right\rangle + b \right] - 1 \right) \qquad (15.3)$$

where α are the Lagrange multipliers. The Lagrangian has to be minimized with respect to w, b and maximized with respect to $\alpha \geq 0$. Eq. (15.3) is then transformed to its dual problem. Hence, the solution of the linearly separable case is given by,

$$\alpha^* = \arg\min_{\alpha} \frac{1}{2} \sum_{i=1}^{L} \sum_{j=1}^{L} \alpha_i \alpha_j y_i y_j \left\langle x_i, x_j \right\rangle - \sum_{k=1}^{L} \alpha_k \qquad (15.4)$$

with constrains, $\alpha_i \geq 0, \quad i = 1, \ldots, L$ and $\sum_{j=1}^{L} \alpha_j y_j = 0$.

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:

$$\alpha^* = \arg\min_{\alpha} \frac{1}{2} \sum_{i=1}^{L} \sum_{j=1}^{L} \alpha_i \alpha_j y_i y_j K(x_i, x_j) - \sum_{k=1}^{L} \alpha_k \qquad (15.5)$$

with constrains, $0 \leq \alpha_i \leq C, \quad i = 1, \ldots, L$ and $\sum_{j=1}^{L} \alpha_j y_j = 0$.

Non-linear mappings (kernel functions) that can be employed are polynomials, radial basis functions and certain sigmoid functions. In this study, polynomial kernel has been employed.

15.3 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 [21, 22]. 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 [21,22]. As demonstrated in [23], the detection of the lip in hue/saturation color is much easier owing to its robustness under a wide range of lip colors and varying illumination condition. Our lipreading database has 22,200 images in total size 64 × 64 pixels from 37 persons. For each person, 60 sequences of images (with ten 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 [24]. 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. Theory of PCA technique for feature extraction can be simply stated as follows. Given a set of N sample images $x_i, i = 1, 2, \ldots, M$ where each image in the set is lithographically re-ordered in L^2 dimensional space and belongs to one of the c classes $\{C_1, C_2, \ldots, C_c\}$. By considering a linear transformation mapping, the original sample in L^2 dimensional space are then transformed into a P-dimensional feature space, where $P \ll M \ll L^2$. The new transformed features $y_i, i = 1, 2, \ldots, M$ is known as subspace and the process of transforming is called projection. In PCA, the transformation process is executed by the following linear transformation:

$$y_i = U^T x_i, \ldots i = 1, 2, \ldots, M \tag{15.6}$$

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

The transformed lip features are then used for the verification process using SVM as classifier. Clean training visual data and clean testing visual data are used for this purpose. In order to model the classifier discriminatively, each speaker model is trained using 3, 6, 10 and 20 client data as well as with 108, 216, 360 and 720 imposter data, respectively. Thus, four types of speaker models are developed for each speaker. During testing, each type of speaker model from each speaker is tested on 40 client data and 1,440 (40 × 36) imposter data from the other 36 persons.

15.4 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 [25, 26]. 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. Let assume the relation between the present sample $x(n)$ and first-order linear combination of the previous p samples as in Eq. (15.7). Consequently, LPC cepstrum coefficients can be derived through the LPC model.

$$x(n) \approx \alpha_1 x(n-1) + \cdots + \alpha_p x(n-p) \tag{15.7}$$

For a time sequence $x(n)$, complex cepstrums $\hat{c}_n$ are represented as below:

$$\hat{c}_1 = -\alpha_1 \tag{15.8}$$

$$\hat{c}_n = -\alpha_n - \sum_{k=1}^{n-1} \left(1 - \frac{k}{n}\right) \alpha_k \hat{c}_{n-k}, \ldots 1 < n \leq p \tag{15.9}$$

$$\hat{c}_n = -\sum_{k=1}^{p} \left(1 - \frac{k}{n}\right) \alpha_k \hat{c}_{n-k}, \ldots n > p \tag{15.10}$$

Experiments for audio only biometric system are divided into two categories. Firstly, to evaluate the system using different numbers of training data and secondly, to evaluate the system based on different SNR levels. Clean training speech data and clean testing speech data are used for this purpose. In order to model the classifier discriminatively, each speaker model is trained using 3, 6, 10 and 20 client data as well as with 108, 216, 360 and 720 imposter data, respectively. Thus, four types of speaker models are developed for each speaker. During testing, each type of speaker model from each speaker is tested on 40 client data and 1,440 (40 × 36) imposter data from the other 36 persons.

Consequently, three experiments are conducted in order to evaluate the system based on different SNR levels. For this purpose, clean data are used for speaker modeling. Each speaker model is trained using 20 client training data and 720 (20 × 36) imposter training data. Three levels of corrupted testing data i.e. 30, 20 and 10 dB SNR data are used. During testing, speaker model from each speaker is tested on 40 client data and 1,440 (40 × 36) imposter data from the other 36 persons for each level of the corrupted signals.

15.5 Fusion System Implementation

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 used to discriminate the incoming speech signal either as high quality or low quality speech signals. Modeling data are taken from the training data set. The indicator system is constructed to differentiate clean speech signal (high quality) from 30, 20 and 10 dB SNR speech signal (low quality). We have used 2,960 training data and 5,920 testing data for this task. This system is capable to achieve 99.66% accuracy.

In fixed weighting approach, the weight to be used for the fusion system is estimated by first running the audio subsystem and visual subsystem separately using the enrollment data. The fusion system is a soft fusion system that uses raw scores from audio and visual subsystems. The scores from each subsystem are then normalized by using min-max normalization technique. In this case, the minimum and maximum scores are transformed to 0 and 1, respectively. The normalized values are then combined by using a weighted summation fusion. 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. The optimum weight is calculated at $w = 0.4$ in this experiment. For the adaptive weighting approach, each audio testing data is first checked for its quality by the audio reliability indicator system. After the speech quality measurement process is completed, the system will decide the weight for the fusion process. If the indicator system determines the current speaker's speech signal as clean speech signal, the optimum weight is employed for the fusion system otherwise the visual only system is executed.

15.6 Results and Discussions

Figure 15.2 shows the performances of the visual only system using 3, 6, 10 and 20 training data. By increasing the numbers of training data to the system, a great improvement in the GAR is observed. At FAR of 0.1%, the GAR of the 3 training data system is 85%. By using 6 and 10 training data, the percentage of GAR increases to 96%. Subsequently, the system reaches to 98% GAR when 20 training data are used. Almost 100% GAR is found for 3, 6, 10 and 20 training data systems when the FAR is equal to 35%, 4%, 3% and 0.2%, respectively. EER performances are observed as 0.27%, 0.94%, 1.15% and 2.70% for system using 20, 10, 6 and 3 training data, respectively.

The audio only system performances using 3, 6, 10 and 20 training data are showed in Fig. 15.3. The increment of the numbers of training data increases the performance of GAR. At FAR of 1%, the GAR of the 3, 6, 10 and 20 training data systems are 83%, 93%, 94% and 96%, respectively. This study observes that the performance of 20 training data system is capable to accomplish almost 100% GAR

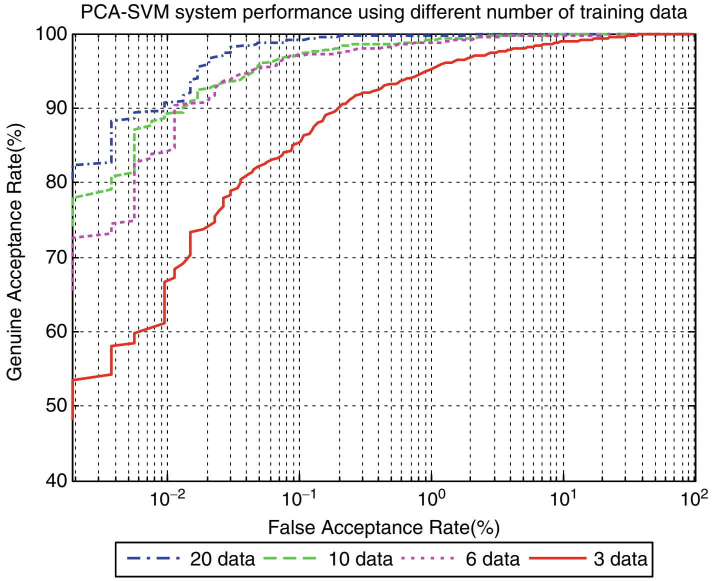

Fig. 15.2 ROC curve for visual only system

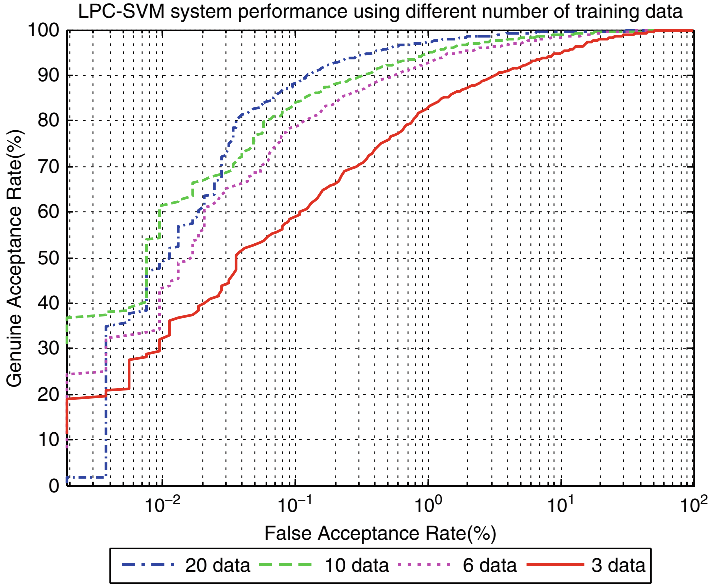

Fig. 15.3 ROC curve for visual only system

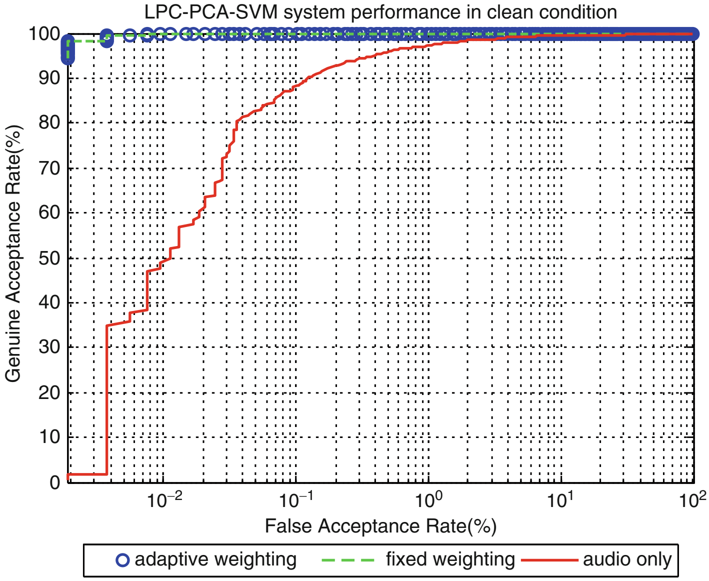

Fig. 15.4 System performances in clean condition

at FAR of 7% compared to FAR of 20%, 25% and 50% for the 10, 6 and 3 training data systems, respectively. System performances based on EER are observed as 1.79%, 2.70%, 3.50% and 6.60% for system using 20, 10, 6 and 3 training data, respectively.

The performance of the adaptive weighting, fixed weighting and audio only system in clean condition are given in Fig. 15.4. The 100% GAR performance is evaluated at FAR is equal to 0.004% for the adaptive weighting and fixed weighting system compared to 35% GAR performance for the audio only system at the same percentage of FAR. On the other hand, it is observed that the audio only system reaches nearly 100% GAR at FAR of 7%. System performances based on EER are observed as 0.067%, 0.067% and 1.79% for adaptive weighting system, fixed weighting system and audio only system, respectively.

The performance of the adaptive weighting, fixed weighting and audio only system based on 30 dB SNR data are given in Fig. 15.5.

Hundred percent of GAR performance is evaluated for the adaptive weighting system at FAR equal to 0.3% compared to FAR equal to 1% for fixed weighting system. The GAR performance at the FAR equal to 0.3% for fixed weighting system is 96% meanwhile GAR performance for the 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 weighting system, fixed weighting system and audio only system, respectively.

The performance of the adaptive weighting, fixed weighting and audio only system based on 20 dB SNR data are given in Fig. 15.6. The 100% GAR is evaluated at

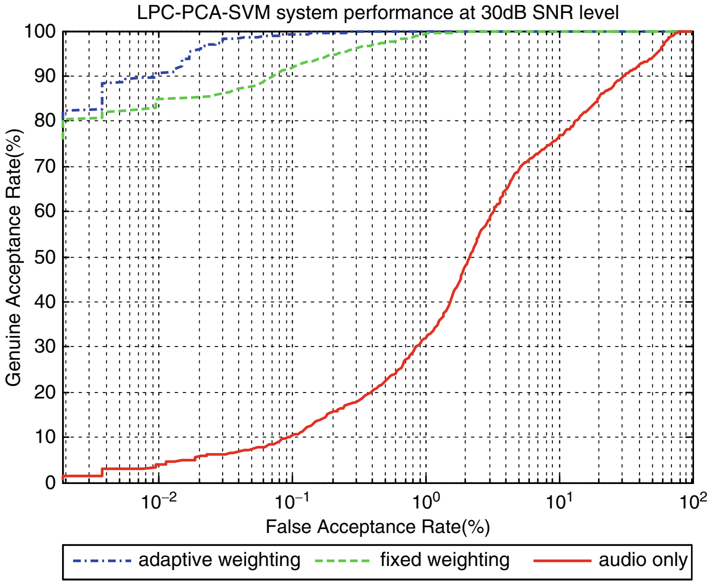

Fig. 15.5 System performances at 30 dB SNR level

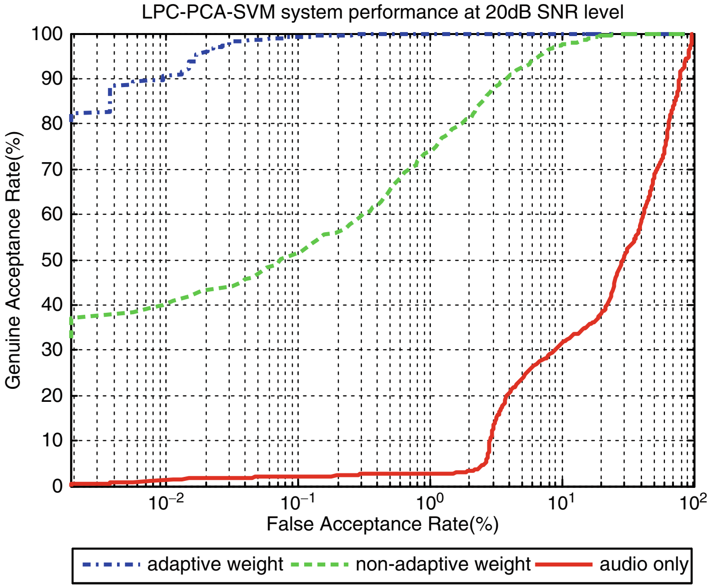

Fig. 15.6 System performances at 20 dB SNR level

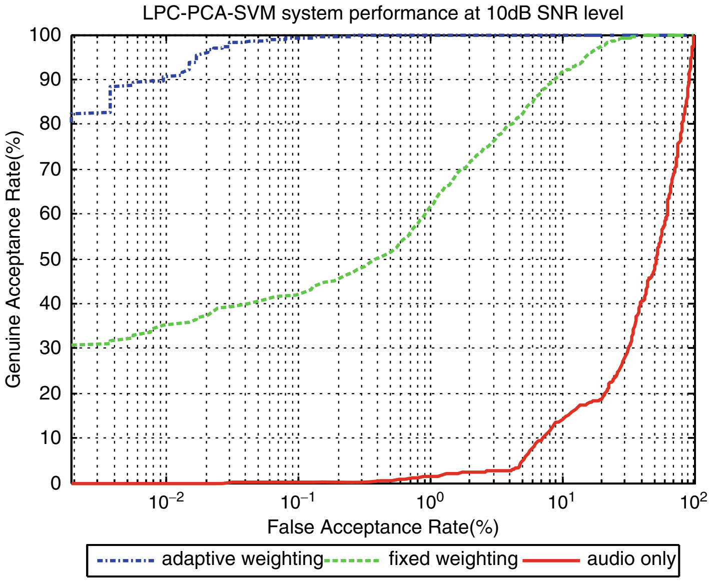

Fig. 15.7 System performances at 10 dB SNR level

FAR is equal to 0.3% for the adaptive weighting system compared to 60% GAR for fixed weighting system at the same FAR percentage. In contrast, the fixed weighting system reaches to 100% GAR at FAR equal to 11%. At the FAR 0.3%, the audio only system simply attains 4% GAR. System performances based on EER are observed as 0.27%, 6% and 40.75% for adaptive weighting system, fixed weighting system and audio only system, respectively.

The performance of the adaptive weighting, fixed weighting and audio only system based on 10 dB SNR data are given in Fig. 15.7.

The 100% GAR performance for adaptive weighting system is evaluated at FAR equal to 0.3% while for fixed weighting system is found at FAR of 35%. In contrast, the performance at FAR of 0.3% for fixed weighting system 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 weighting system, fixed weighting system and audio only system, respectively.

15.7 Conclusions

The performances of the adaptive weighting, fixed weighting and audio only system at different SNR levels have been reported for comparison. This study proved that the proposed SVM indicator system is viable for estimating the quality of speech signal and the implementation of the adaptive weighting approach is imperative for uncertain audio condition. The advantage of using the adaptive weighting instead of fixed weighting is to avoid unreliable scores to be fused together in fusion systems that can spoil the accuracy of the total scores. 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 audio indicator system and visual verification system. Future work will be devoted on all SNR levels and different types of noises. Noise eradication techniques will also be experimented to the audio and visual subsystems so as to enhance the system performance.

References

1. Campbell, J.P.: Speaker recognition: a tutorial. Proc. IEEE **85**, 1437–1462 (1997)
2. Reynolds, D.A.: An overview of automatic speaker recognition technology. Proc. IEEE Acoustics Speech Signal Processing **4**, 4072–4075 (2002)
3. Ramli, D.A., Samad, S.A., Hussain, A.: In: Corchado, E., et al. (ed.) Score Information Decision Fusion using Support Vector Machine for a Correlation Filter Based Speaker Authentication System, vol **53**, pp. 235–242. Springer, Berlin, Heidelberg (2008)
4. Brunelli, R., Falavigna, D., Stringa, L., Poggio, T.: Automatic person recognition by using acoustic and geometric. Mach. Vis. Appl. **8**, 317–325 (1995)
5. Brunelli, R., Falavigna, D.: Personal identification using multiple cue. IEEE Trans. Pattern Anal. Mach. Int. **17**(3), 955–966 (1995)
6. Dieckmann, U., Plankensteiner, P., Wagner, T.: SESAM: A biometric person identification system using sensor. Pattern Recog. Lett. **18**(9), 827–833 (1997)
7. Jourlin, P., Luettin, J., Genoud, D., Wassner, H.: Integrating acoustic and labial information for speaker identification and verification. Proc. 5th European Conf. Speech, Commun. Technol. **3**, 1603–1606 (1997)
8. Sanderson, C., Paliwal, K.K.: Multi-modal person verification system based on face profile and speech. Fifth International Symposium on Signal Processing and Its Applications, pp. 947–950 (1999)
9. Pan, H., Liang, Z.P., Huang, T.S.: Fusing audio and visual features of speech. Proc. IEEE Int. Conf. Image Processing **3**, 214–217 (2000)
10. Chu, S.M., Marcheret, V.L.E., Neti, C., Potamianos, G.: Multistage information fusion for audio-visual speech recognition. Proc. IEEE Int. Conf. Multimedia Expo, pp. 1651–1654 (2004)
11. Gurban, M., Thiran, J.P.: Using entropy as a stream reliability estimate for audio-visual speech. 16th European Signal Processing Conference (2008, in press)
12. Potamianos, G., Neti, C.: Stream confidence estimation for audio-visual speech. Proc. Int. Conf. Spoken Language **3**, 746–749 (2000)
13. Heckmann, M., Berthommier, F., Kroschel, K.: Noise adaptive stream weighting in audio-visual speech. EURASIP J. Appl. Signal Process. **2002**(11), 1260–1273 (2002)
14. Chetty, G., Wagner, M.: Robust face-voice based speaker verification using multilevel. Image Vision Comput. **26**(9), 1249–1260 (2008)
15. Wark, T., Sridharan, S.: A syntactic approach to automatic lip feature extraction for speaker identification. IEEE Int. Conf. Acoustics Speech Signal Processing **6**, 3693–3696 (1998)

16. Broun, C.C., Zhang, X., Mersereau, R.M., Clements, M.: Automatic speechreading with application to speaker verification. IEEE Int. Conf. Acoustics Speech Signal Processing **1**, 685–688 (2002)
17. Fox, N.A., Reilly, R.B.: Robust multi-modal person identification with tolerance of facial expression. Proc. IEEE Int. Conf. System Man Cybernetics **1**, 580–585 (2004)
18. Sanderson, C., Paliwal, K.K.: Noise compensation in a multi-modal verification system. Proc. Int. Conf. Acoustics, Speech Signal Processing **1**, 157–160 (2001)
19. Gunn, S.R.: Support vector machine for classification and regression. Technical Report, University of Southampton (2005)
20. Wan, V., Campbell, W.M.: Support vector machines for speaker verification and identification. Proc. Neural Networks Signal Processing **2**, 775–784 (2000)
21. Chetty, G., Wagner, M.: Liveness verification in audio-video speaker authentication. Proc. Int. Conf. Spoken Language Processing ICSLP 04, pp. 2509–2512 (2004)
22. Chetty, G., Wagner, M.: Automated lip feature extraction for liveness verification in audio-video authentication. Proc. Image Vision Comput., pp. 17–22 (2004)
23. Matthews, I., Cootes, J., Bangham, J., Cox, S., Harvey, R.: Extraction of visual features for lipreading. IEEE Trans. Pattern Anal. Mach. Intell. **24**(2), 198–213 (2002)
24. Kirby, M., Sirovich, L.: Application of the Karhunen-Loeve procedure for the characterisation of human. IEEE Trans. Pattern Anal. Mach. **12**(1), 103–108 (1990)
25. Rabiner, L.R., Juang, B.H.: Fundamental of Speech Recognition. Prentice-Hall, New York (1993)
26. Furui, S.: Cepstral analysis technique for automatic speaker verification. IEEE Trans. Acoust Speech Signal Process. **29**(2), 254–272 (1981)