World Applied Sciences Journal 12 (11): 2160-2167, 2011
ISSN 1818-4952

Performances of Weighted Sum-Rule Fusion Scheme in Multi-Instance and Multi-Modal Biometric Systems

[1]Dzati Athiar Ramli, [1]Nurul Hayati Che Rani and [2]Khairul Anuar Ishak

[1]School of Electrical & Electronic Engineering, USM Engineering Campus,
Universiti Sains Malaysia, 14300, Nibong Tebal, Pulau Pinang, Malaysia
[2]Faculty of Engineering and Built Environment,
Universiti kebangsaan Malaysia, 43600, Bangi, Selangor, Malaysia

Abstract: Biometric speaker authentication systems use speaker specific information in speech signal to discriminate individuals. However, intra-speaker variations such as the changing of speaking rates, emotional and health conditions can affect system performances. One of the solutions to this limitation is to implement multibiometric system approach by combining different sources of biometric information. In this study, two approaches of m ultibiometric systems i.e. multi-instance systems and multi-modal system are experimented. Multi-instance systems consider a combination of information from several samples of different verbal extracted from the same modality while multi-modal systems is the fusion of different information extracted from different modality. Due to the different performances of each instance and each modality in the respective multi-instance system and multi-modal system, the use of weighted sum-rule fusion is suggested in this study. Performances based on sum-rule fusion are also evaluated for comparison. This research focuses on the score level fusion and Min-Max normalization technique is employed for score normalization. For s peech signal feature extraction, the information in term of Mel Frequency Cepstral Coefficient (MFCC) is extracted while region of interest (ROI) of face images has been used as a second modality for the multi-modal systems. The Support Vector Machine (SVM) classifier is executed for the verification process. Experimental results prove that performances of multi-modal systems with weighted sum-rule fusion are outstanding compared to the other system performances. EER performances for multi-modal system with w eighted sum-rule fusion, multi-modal system with sum-rule fusion, multi-instance systems with weighted sum-rule fusion, multi-instance systems with sum-rule fusion and speech signal single system (verbal zero) are observed as 0.0563%, 0.2778%, 1.9904%, 2.0261% and 4.3206%, respectively.

Key words: Multi-modal · Multi-instance · sum-rule and weighted sum-rule · Speech signal biometrics

INTRODUCTION

According to Reynolds 2002 [1], biometrics is a technology used for authentication and identification purposes by measuring physical or behavioral information of an individual. In recent years, the use of biometric data in recognition system has been widely implemented in many applications such as in access control system, banking system, computer network security and law enforcement. The use of biometric data for authenticating and identifying person is more secure compared to the traditional ways which requires key, smart card and password (PIN). This is because key, smart card and password can be easily stolen, duplicated, forgotten or guessed by the imposter. Biometric traits can be classified as physiological and behavioral categories. Face, fingerprint, finger vein, iris pattern, palm print, hand geometry and ear shape are examples of physiological traits meanwhile gait and signature are known as behavioral trait. Speech can be considered either as physiological or behavioral trait because both physical information such as pitch and nasality and behavioral information for examples pronunciation and conversational styles are contained in speech [2].

Corresponding Author: Dzati Athiar Ramli, School of Electrical & Electronic Engineering, USM Engineering Campus, Universiti Sains Malaysia, 14300, Nibong Tebal, Pulau Pinang, Malaysia.
Tel: 04-599599, Ext: 6028, E-mail: dzati@eng.usm.my.

This study concentrates on the implementation of biometric speaker authentication systems. This system uses the specific information in speech signal which related to the anatomical distinctiveness and speaking tendency to discriminate speakers. These systems verify the claimant by accepting the genuine user whilst rejecting the imposter. In order to be correctly classified, the extracted speech information should have high inter speaker variations and low intra-speaker variations.

The structure of speaker verification systems consists of four main components i.e. data acquisition, feature extraction, pattern-matching and decision [3]. Data acquisition is divided into two tasks i.e. enrollment data acquisition and current speaker data acquisition. Enrollment data acquisition is a process to acquire training speech data from registered speakers for speaker's model or speaker's template. Registered speakers are authorized users who have legal permission to access the systems. Consequently, the current speaker data acquisition is a task to obtain the current input speech signal by the claimant or user of the systems for verification.

In the feature extraction process, the most relevant information from the speech sample is extracted to form a feature vector. Most commonly, spectral based features either using LPC or FFT analysis are used in most speaker verification systems [1]. High level feature based approaches for instances prosodic dynamics, pitch gestures and phone streams are also addressed in Campbell *et al.* (2003) [4]. The feature vector from each registered speaker obtained during enrollment is stored in the database as the speaker's model or the speaker's template. For verification purpose, a similar feature extraction scheme must be executed for the current speaker's speech signal. The pattern-matching component involves the procedure to verify the current speaker by comparing current speaker's feature set from its corresponding model/template in the database. Pattern-matching techniques such as DTW, VQ, ANN, HMM, GMM and SVM are commonly used as classifier for this task. In the decision component, the current speaker's match score is then compared to a threshold that earlier specified by the systems and a decision whether to accept or reject the current speaker is then made in this step.

According to Reynolds [1], the advantages of employing speech signal modality in biometric systems are due to its natural source, effortless generated, require a little custom hardware for data collection, low computation requirement for feature extraction and matching and also high accuracy performances. However, as reported in Jain *et al.* (2004) [5], the use of single trait biometric has several limitations and the systems tend to obtain low performance especially in noisy environment. This becomes the main problem for utilizing speech signals for biometric systems. There are many factors contribute to the verification error for instances the change of speaking rates, sickness (e.g. influenza and cough), extreme emotional state (e.g. anger and depression) and aging. Additionally, different microphones and channels also affect the accuracy of the system performance. So that, the implementation of biometric systems has to properly distinguish 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.

By combining different sources of biometric data some of these setbacks can be overcome as reported by several researches for instances in [6-9]. Incorporating fusion technique in non-biometric systems has also been reported in many studies recently. The performances of web ranking using the combination of content and context features have been experimented in [10]. The integration of global positioning system and inertial navigation system for accurate navigation technology has been found in [11]. According to Ross and Jain (2007) [12], multibiometric systems are presented by multiple sources of biometric information involving the use of biometric fusion. Five categories of multibiometric systems have been reported in literatures i.e. multi-sensorial systems, multi-instance systems, multi-modal systems, multi-algorithmic systems and multi-sample systems. The level of fusion in multibiometric systems is categorized into two categories i.e. fusion before matching and fusion after matching. Fusion before matching includes fusion at the sensor and feature levels whereas fusion after matching includes fusion at the score and decision levels.

Fixed weighting approach can be found in Brunelli *et al.* (1995) [13]. This study has presented fusion of scores produced independently by speaker recognition system and face recognition system using a weighted merged score.

The optimal weight, w_{opt} is found by maximizing the performance of the integrated system on one of the available training set. The identification of 51% is achieved for speech alone system and 92% for the face alone system. The performance of the integration system using the optimal weight is observed up to 95%. In another experiment, Brunelli and Falavigna (1995) [14], have evaluated a weighted product approach to fuse two voice features i.e. static and dynamic and three face features i.e. eye, noise and mouth. The tan-estimator is used for score normalization and weighted geometric

average is 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.

Dickmann *et al.* (1997) [15] have combined different biometric cues i.e. voice and face image. The recognition decision is based on total weighted activations, the EER performance of face recognition, voice recognition and integrated face and voice recognition with w = 0.2 are obtained as 3%, 3.4% and 1.5% from this experiment. The reduction of the EER values for the integrated face and voice system shows that this system outperforms the performance of both single biometric systems. Jourlin *et al.* (1997) [16] have integrated the scores of speech modality and lip modality using weighted summation fusion. The performance of the integrated system outperforms the performance of each subsystem and reduces the false acceptance rate of the speech subsystem from 2.3% to 0.5%.

In this study, two types of multibiometric systems i.e. multi-instance systems and multi-modal systems are developed. For multi-instance systems, scores from several samples of different verbal i.e. zero, seven and eight are combined. Whereas, for the multi-modal systems, the verification scores from speech and face subsystem are fused together. This study concentrates on score level fusion and due to the different performances of each instance and each modality in the respective multi-instance system and multi-modal system; the use of weighted sum-rule fusion is then suggested in this study. Performances based on sum-rule fusion are also evaluated for comparison. Then, normalization methods such as min-max normalization, z-score normalization and tanh-estimators normalization need to be applied in order to normalize the scores. For simplicity, min-max normalization technique is executed in this research.

The first objective of this study is to develop speech signal single systems for speaker verification. Then, the second objective is to develop multi-instance and multi-modal systems with sum-rule and weighted sum-rule fusion for speaker verification. Finally, the performances of the developed systems are compared for their achievements.

Methodology: In this study, both audio and visual data are obtained from Audio- Visual Digit Database (Sanderson and Paliwal 2001) [17]. This database contains digitized audio signals which monophonic 16 bit, 32 kHz and in WAV format corresponding to the recording voices of 37 speakers (16 female and 21 male). The recording is done in three different sessions. In each session, each speaker performed 20 repetitions of digit zero to nine hence 60 audio data for each speaker from all sessions so that it consist of 2220 data for each digit. In total 6660 audio data from entire speakers have been used for this study which is digit zero, seven and eight. The visual data of 37 speakers is stored as a sequence of JPEG images with a resolution of 512 × 384 pixels. For each speaker, 60 sequences of frontal face images (20 sequences from each session) hence in total of 2220 images from entire speakers. For the purpose of this study, the first session are used for speaker's model while data from the second and third session are used as testing data.

For the purpose of this study, Support Vector Machine (SVM) is used as a classifier to execute the pattern matching phase. SVM classifier is a discriminative model which involves data samples for both clients and imposters during training session in order to create a decision boundary. Score level fusion scheme is then considered for the combination of scores from all subsystems. The scores from each subsystem are normalized before fusion using min-max normalization techniques. Normalization is necessary to transform the scores of different ranges to a common range before fusions are conducted. In this case, the minimum and maximum scores are transformed to 0 and 1, respectively. The min-max normalized score, $\breve{s}_i$ from any biometric expert is given by

$$\breve{s} = \frac{s_i - min_{i=1}^{k} s_i}{max_{i=1}^{k} - min_{i=1}^{k}} \quad (1)$$

Where s_i denote the ith match score output and K is the number of the match scores available in the set [18].

In this study, two categories of fusion schemes are used which are sum-rule fusion scheme and weighted sum-rule fusion scheme. The method of sum-rule based fusion is stated in the equation (2). From the set of normalized scores of all speakers, the fused score *fs* is calculated using the equation

$$f_s = w_1 x_1 + ... + w_m x_m \quad (2)$$

The s_i refers to weight which is allocated to the sample verbal-*i* for multi-instance system and modality-*i* for multi-modal system, for i=1,….,m.

For the sum-rule based, equal weights are used for each sample verbal and modality in these experiments. Then equation (2) is simplified to

$$f_s = x_1 + ... + x_m \quad (3)$$

The procedure of weighted sum-rule fusion can be found in Jain and Ross (2004) [19]. From the equation (1), the weights, to ` are varied over the range 0 to 1, such that the limitation $w_1x_1 + ... + w_m x_m =$ is satisfied. Hence, the equation (2) based on three sample verbal in multi-instance system is simplified to

$$f_s = w_1x_1 + w_2x_2 + w_3x_3 \quad (4)$$

Whereas for multi-modal system, the equation (2) based on two modalities is simplified to

$$f_s = w_1x_1 + w_2x_2 \quad (5)$$

Scheidat *et al.* (2007) [20] suggested a method to determine the weights based on the performance of the single system by utilizing the equal error rate (EER) value. The weights of each subsystem are calculated by dividing the single EER with sum of all EERs as shown in equation (6).

$$w_i = \frac{eer_i}{\Sigma_{m=1}^{n} eer_m} \quad (6)$$

A property of this weighting scheme is that the system which obtained the highest EER is multiplied with the smallest weight and vice versa.

Multi-instance Systems: For multi-instance systems, an experiment is conducted by combining three audio subsystems from different models of verbal zero, seven and eight. For modeling purpose, each speaker model is trained using 3 and 6 client training data and 108 and 216 imposter training data, correspondingly. During testing, speaker model from each speaker for each different verbal audio system is tested on 40 client data and 1440 (40x36) imposter data from the other 36 persons. Thus, 1480 scores of testing data are obtained for each verbal audio system.

The scores from each verbal of audio systems are then fused using equation (3) for sum-rule fusion and equation (4) for weighted sum-rule fusion. The weight in the weighted sum-rule fusion scheme is defined from equation (6). Hence, two types of multi-instance systems have been developed in this experiment namely multi-instance system with sum-rule fusion and multi-instance system with weighted sum-rule fusion.

Multi-Modal Systems: For multi-modal systems, two subsystems based on the previous multi-instance and visual subsystems are combined. Two types of multi-instance subsystems are employed i.e. multi-instance subsystem with sum-rule fusion and multi-instance subsystem with weighted sum-rule fusion. For each multi-instance sub-system, each speaker model is trained using 3 client training data and 108 imposter training data.

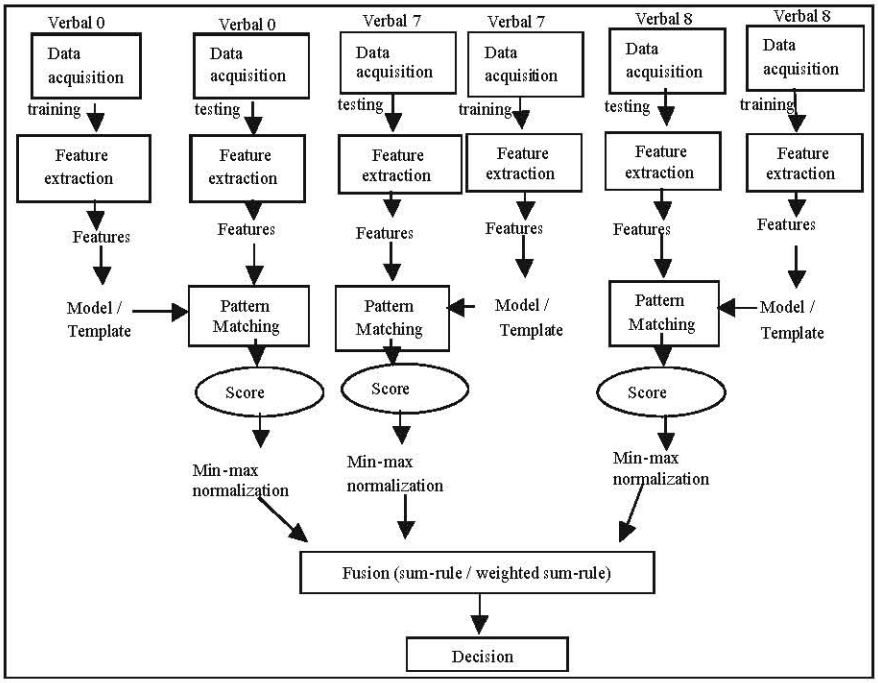

Fig. 1: Architecture of multi-instance system based on different verbal of speech signal

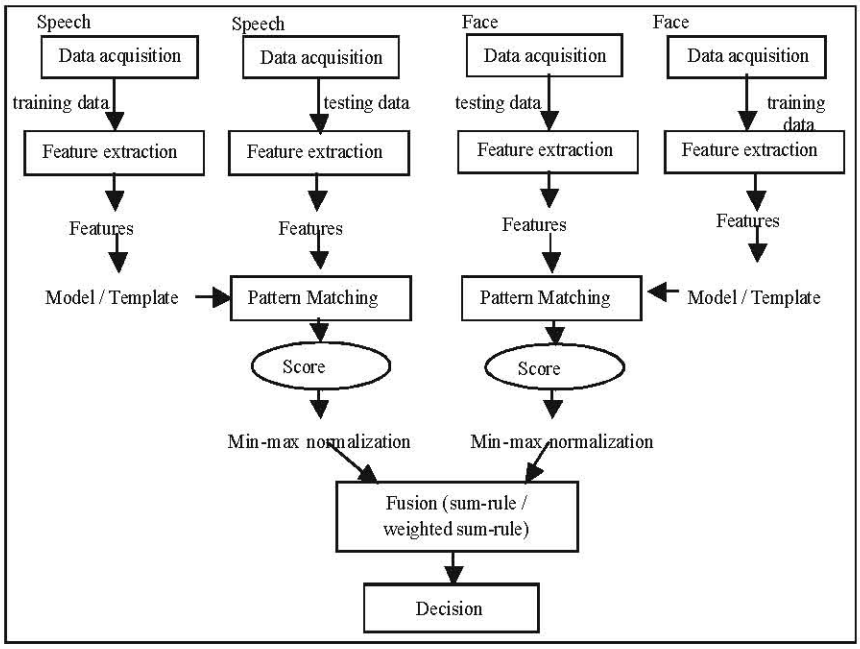

Fig. 2: Architecture of multi-modal systems

During testing, speaker model from each speaker of each multi-instance subsystem is tested with 40 client data and 1440 imposter data from other 36 persons. Thus 1480 scores are obtained for each multi-instance subsystem.

For visual subsystem, each speaker is trained using 3 client data as well as with 108 imposter data. During testing, speaker model from each speaker is tested on 40 client data and 1440 imposter data from other 36 person. Thus 1480 scores are obtained for visual subsystem. The scores from each subsystem are then normalized by using min-max normalization technique. The normalized scores from multi-instance subsystem with sum-rule fusion and visual subsystem are then fused using equation (3) for sum-rule fusion and equation (5) for weighted sum-rule fusion. Similarly, for the normalized scores of multi-instance subsystem with weighted sum-rule fusion and visual subsystem are fused using equation (4) for sum-rule fusion and equation (5) for weighted sum-rule fusion. The weight for weighted sum-rule fusion is defined from equation (6).

RESULTS AND DISCUSSIONS

The performance of a biometric system is significant to determine achievement of the system and it is normally evaluated in terms of false rejection rate (FRR) and the false acceptance rate (FAR). FRR also known as false match rate i.e. when the authorized persons are rejected as being an impostor. In term of sensitivity or genuine acceptance rate (GAR), it can be explained as the percentage of authorized persons is admitted by the system. Equal Error Rate (EER) is another method to measure the performance of biometric systems. EER is the equal value of FAR and FRR. In general, the lower EER value, the higher the performance of the system. A receiver operating characteristics (ROC) graph can be used to interpret the system performance of the systems details explanation on ROC can be found in Fawcett (2006) [21].

Figure 3 compares the performances of single modal biometric systems, multi-instance system and multi-modal system based on 3 training data. For single modal speech biometric systems, three systems are developed which employed verbal zero, verbal seven and verbal eight. Multi-instance system is developed based on sum-rule fusion scheme. The multi-modal biometric system is constructed by integrates the multi-instance (sum-rule) speech system and single modal face system. A sum-rule fusion scheme has been used in this multi-modal biometric system. At FAR of 1%, the GAR percentages of single modal of verbal zero, verbal seven and verbal eight are observed as 88%, 74% and 72%, respectively whereas for multi-instance (sum-rule) system and multi-modal (sum-rule) are found as 96% and 100%, respectively. The system performances based on EER are shown in Table 1.

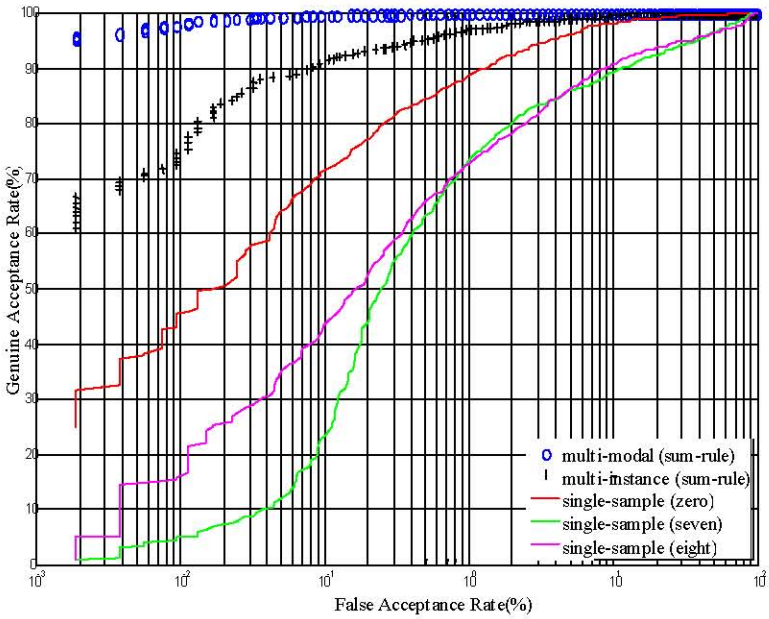

Fig. 3: Performances of single modal, multi-instance (sum- rule) system and multi-modal biometric system (sum-rule) based on 3 training data

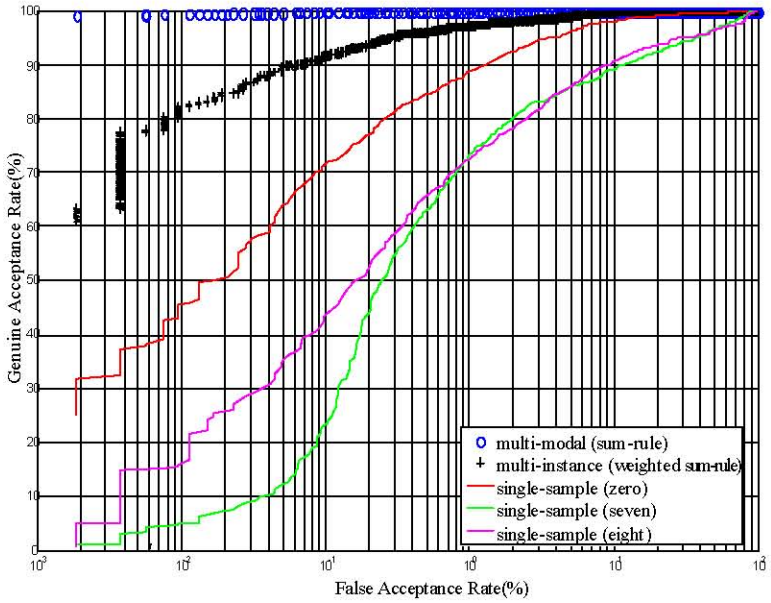

Fig. 4: Performances of single modal, multi-instance (weighted sum-rule) system and multi-modal biometric system (sum-rule) based on 3 training data

Consequently, Figure 4 compares the performances of single modal biometric system, multi-instance system and multi-modal biometric system based on 3 training data. For single modal speech biometric systems, three systems are developed which employed verbal zero, verbal seven and verbal eight. Multi-instance system is developed based on weighted sum-rule fusion scheme. The multi-modal biometric system is constructed by integrates the multi-instance (weighted sum-rule) speech system and single modal face system. A sum-rule fusion scheme has been used in this multi-modal biometric system. At FAR of 1%, the GAR of single modal system of verbal zero, verbal seven and verbal eight are 88%, 74% and 72%, respectively whereas for multi-instance (weighted sum-rule) system and multi-modal (sum-rule) system, the GAR percentages are 97% and 100%, respectively. The system performances based on EER are shown in Table 2.

Table 1: Performances of single modal, multi-instance (sum-rule) system and multi- modal biometric (sum-rule) based on 3 training data

System	Multi-modal	Multi-instance	Single-modal (zero)	Single-modal (seven)	Single-modal (eight)
EER	0.2778	2.0261	4.3206	10.4148	9.6181

Table 2: Performances of single modal, multi-instance (weighted sum-rule) system and multi-modal biometric (sum-rule) based on 3 training data

System	Multi-modal	Multi-instance	Single-modal (zero)	Single-modal (seven)	Single-modal (eight)
EER	0.1361	1.9904	4.3206	10.4148	9.6181

Table 3: Performances of single modal, multi-instance (weighted sum-rule) system and multi-modal biometric (weighted sum-rule) based on 3 training data

System	Multi-modal	Multi-instance	Single-modal (zero)	Single-modal (seven)	Single-modal (eight)
EER	0.0563	1.9904	4.3206	10.4148	9.6181

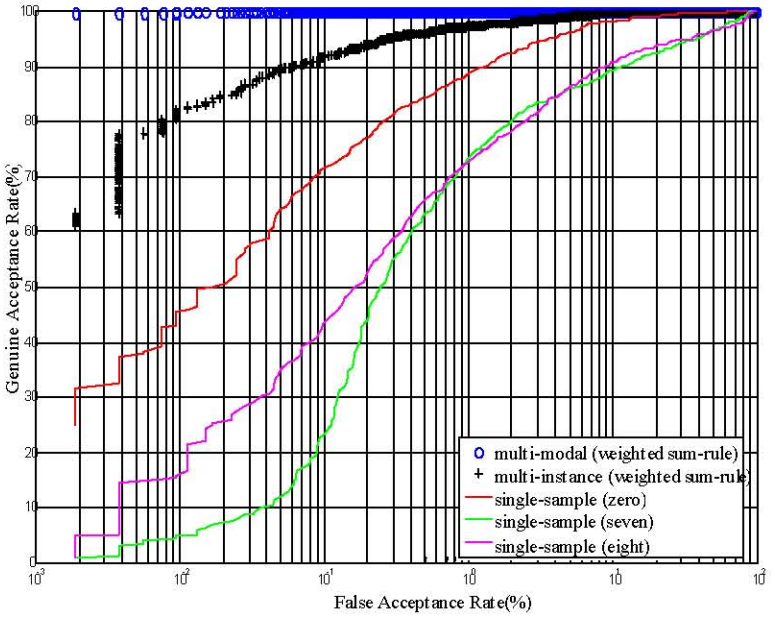

Fig. 5: Performances of single modal, multi-instance (weighted sum-rule) system and multi-modal biometric system (weighted sum-rule) based on 3 training data.

Figure 5 compares the performances of single modal systems, multi-instance system and multi-modal biometric systems based on 3 training data. For single modal speech biometric systems, three systems are developed which employed verbal zero, verbal seven and verbal eight. Multi-instance system is developed based on weighted sum-rule fusion scheme. The multi-modal biometric system is constructed by integrating the multi-instance (weighted sum-rule) speech system and single modal face system. A weighted sum-rule fusion scheme has been used in this multi-modal biometric system. At FAR of 1%, the GAR percentages of single modal system of verbal zero, verbal seven and verbal eight are 88%, 74% and 72%, respectively whereas for multi-instance (weighted sum-rule) and multi-modal (weighted sum-rule) systems are given as 97% and 100% respectively. In general, this study concludes that the multi-modal (weighted sum-rule) system outperforms the others. The system performances based on EER are shown in Table 3.

CONCLUSIONS

The performances of three types of biometric systems namely speech signal single system, multi-instance speech system and multi-modal system were evaluated in this study. Implementation of multibiometric system approaches i.e. multi-instance and multi-modal systems enhance the single system performances. Furthermore, execution of weighted sum-rule fusion is imperative in boosting the system performances due to the differences of each instance and modality performances in the respective multi-instance and multi-modal systems.

This study also evaluated that the performances of multi-modal systems are more outstanding compared to the multi-instance systems.

ACKNOWLEDGMENT

This research is supported by the following research grants: Research University Grant, Universiti Sains Malaysia, 1001/PELECT/814098 and Incentive Grant, Universiti Sains Malaysia.

REFERENCES

1. Reynolds, D.A., 2002. An overview of automatic speaker recognition technology. IEEE Transactions on Acoustics, Speech and Signal Processing, 4: 4072-4075.
2. Jain, A.K., R. Bolle, and S. Pankanti, 1999a. *Biometrics: Personal identification in networked society*. New York: Springer Science & Business media, Inc.
3. Campbell, J.P., 1997. Speaker Recognition: A tutorial. Proceeding of IEEE, 85(9): 1437-1462.
4. Campbell, W.M., 2003. A SVM/HMM system for speaker recognition. IEEE ICASSP, 2: 209-212.
5. Jain, A.K., A. Ross and S. Prabhakar, 2004. An Introduction to Biometric Recognition, IEEE transactions On circuits And Systems For Video Technol., 14(1).
6. 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.
7. 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, pp: 229-232.
8. Samad, S.A., D.A. Ramli and A. Hussain, 2007. A Multi-Sample Single-Source Model using Spectrographic Features for Biometric Authentication, IEEE International Conference on Information, Communications and Signal Processing, CD ROM.
9. Ramli, D.A., S.A. Samad and A. Hussain, 2010. A Correlation Filter Based Biometric S peaker Authentication Systems, World Appl. Sci. J., 9(3): 259-267.
10. Keyhanipour, A.H., M. Piroozmand and K. Badie, 2009. A Neural Framework for Web Ranking Using Combination of Content and context Features, World Applied Sci. J., 6(1): 6-15.
11. Hassan, A.M. and S. Khairulmizam, 2009. Integration of Global Positioning System and Inertial Navigation System with Different Sampling Rate using Adaptive Neuro Fuzzy Inference System, World Applied Sciences J., 7. (Special Issue of Computer & IT): 98-106.
12. Ross, A. and A.K. Jain, 2007. Fusion Techniques in multibiometric systems. In H a mmoud, R.I., Abidi, B.R. & Abidi, M.A. Face Biometrics for Personal Identification, pp: 185-212. Berlin: Springer-Verlag Inc.
13. Brunelli, R., D. Falavigna, T. Poggio and L. Stringa, 1995. Automatic Person Recognition by Using Acoustic and Geometric. Machine Vision & Applications, 8: 317-325.
14. Brunelli, R. and D. Falavigna, 1995. Person identification using multiple cues. IEEE Transactions on Pattern Analysis and Machine Intelligent, 17(10): 955-966.
15. Dickmann, U., P. Plankensteiner and T. Wagner, 1997. SESAM: A biometric person identification system using sensor fusion. Pattern Recognition Lett., 18(9): 827-833.
16. Jourlin, P., J. Luettin, D. Genoud and H. Wassner, 1997. Integrating Acoustic and Labial Information for Speaker Identification and Verification. Proceeeding 5th European Conference Speech Communication and Technol., (EUROSPEECH) 3: 1603-1606.
17. Sanderson, C. and K.K. Paliwal, 2001. Noise Compensation in a Multi-Modal Verification System, Proceeding of International Conference on Acoustics, Speech and Signal Processing, pp: 157-160.
18. Jain, A., K. Nandakumar and A. Ross, 2005. Score normalization in multimodal biometric systems, Pattern Recognition, 38: 2270-2285.
19. Jain, A.K. and A. Ross, 2004. Multibiometric systems. Communications of the ACM. Special Issue on Multimodal, 47(1): 34-40.
20. Scheidat, T., C. Vielhauer and J. Dittman, 2007. Single-semantic MultiInstance fusion of handwriting based biometric Authentication systems.IEEE 1-4244-1437-7/07 ICIP 2007.
21. Fawcett, T., 2006. An Introduction to ROC analysis.Pattern Recognition Lett., 2 7: 861-874.