World Applied Sciences Journal 10 (7): 847-852, 2010
ISSN 1818-4952

A Multi-Instance Speech Signal Data Fusion Approach for Biometric Speaker Authentication System Enhancement

[1]*Dzati Athiar Ramli,* [2]*NurulHayati Che Rani and* [3]*Khairul Anuar Ishak*

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

Abstract: In this study, an alternative approach for combining biometric data to enhance the performance of biometric system is proposed. This approach namely multi-instance data fusion scheme which combines speech signal data with different verbal as features to the biometric systems. For speech signal feature extraction, the information in term of Mel Frequency Cepstral Coefficient (MFCC) is extracted while Support Vector Machine (SVM) classifier is used for pattern matching. The Min-Max normalization technique is employed to normalize the score of each verbal classification. Then, the sum-rule fusion scheme is executed before accepting or rejecting the client at the decision stage. The main objective of this study is to evaluate the performances of the multi-instance fusion scheme by comparing this approach to the single-sample based systems. Experimental results show that the proposed approach can be a viable fusion scheme. The EER performances of the multi-instance data fusion are observed as 2.0261% compared to the EER performances of 4.3206%, 9.6181% and 10.4148% for three different single-sample based systems. Apart from this improvement, the multi-instance approach can reduce the number of training data used in feature modeling whilst maintaining the similar performances.

Key words: Multi-instance · Single-sample · Data fusion · Speech signal biometrics

INTRODUCTION

The execution of a single source of biometric information in biometric systems suffers from limitations such as noise data, high intra-class variations, non-universality and spoof attacks [1]. Due to these problems, the systems may not be able to achieve desired performances in the real applications. By combining different sources of biometric data some of these setbacks can be overcome as reported by several researches for instances in [2-6]. 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 [7]. The integration of global positioning system and inertial navigation system for accurate navigation technology has been found in [8]. In this research, an adaptive neuro fuzzy inference system has been used for the fusion scheme [8].

This paper focuses on multi-instance data fusion approach. Multi-instance systems can be defined as the combination of the biometric information extracted from different sources of the same biometric modality. The implementation of multi-instance systems has been found in [9, 10]. Prabhakar & Jain[9] uses the left and right index finger while Jang *et al.* [10] employs the left and right iris as sources of information. Extending this notion, this research proposes the combination of different verbal of speech signal modality as information to the biometric systems.

The audio front end module and the verification module have been developed in this investigation. In the audio front-end module, the information in term of Mel Frequency Cepstral Coefficient (MFCC) is extracted from the raw speech data from the subjects in the database. In the verification module, single-sample system and multi-instance system have been developed based on Support Vector Machine (SVM) classifier. Three different verbal, i.e zero, seven and eight have been used as data

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

modeling in single-sample system. For the multi-instance system, the fusion of three verbal scores is done using the sum-rule scheme. The performances of the single-sample systems and multi-instance system are then evaluated in term of the Equal Error Rate (EER), Genuine Acceptance Rate (GAR) and False Acceptance Rate (FAR).

The database used in this study is the Audio-Visual Digit Database [2]. The database consists of video and the corresponding audio of people reciting digits zero to nine. The video of each person is stored as a sequence of JPEG images with a resolution of 512 x 384 pixels while the corresponding audio provided is a monophonic, 16 bit, 32 kHz, WAV format.

Audio Front End Module: Software programming is executed using MATLAB version 7.0 (release 14) and signal processing toolbox, image processing toolbox and Voicebox are utilized. In data acquisition, voice which is a pressure wave is converted into numerical values in order to be digitally processed in feature extraction. For this purpose, a microphone is used to allow the pressure sound wave to be converted into electrical signal. This continuous electrical signal is then transformed using a sampler and A/D converter into a digital signal. This process is commonly referred as digitization which consists of sampling, quantization and coding.

Audio feature extraction consists of pre-emphasis, framing, windowing and parameter analysis. This study implements MFCC processing for the parameter analysis as described in [11]. A pre-emphasis of high frequencies is required to compress the signal dynamic range by flattening the spectral tilt in order to raise the SNR. Then, due to spectral evaluation is reliable for a stationary signal whose characteristic are invariant with respect to time, short time analysis is performed by framing the pre-emphasized signal. 15-30 ms duration for each frame with 50% overlapping has been used for this purpose. Consequently, the use of window function is important to minimize the signal discontinuities at the beginning and end of each frame by zeroing out the signal outside the region of interest. The Hamming window has been employed in this study.

Spectral analysis returning Mel Frequency Cepstral Coefficient (MFCC) is processed on the Fourier transform. Computing the Discrete Fourier Transform (DFT) of all frames of the signal is the first step in MFCC processing. The result obtained after this step is referred as signal's spectrum. The second step is a filter bank processing. Spectral features are generally obtained as the exit of filter banks which properly integrate a spectrum at defined frequency. The third step is the log energy computation which consists of computing the logarithm of the square magnitude of the filter bank outputs. The final procedure for MFCC processing is mel frequency cepstrum computation that performs the inverse DFT on the logarithm of the magnitude of the filter bank output. Further improvement in performance is achieved by considering the relevant information from the dynamic evolution of the speech signal or delta coefficients. The overall data acquisition and feature extraction process is summarized in Figure 1.

Verification Module: System verification has been designed based on support vector machine (SVM) classifier which requires the development of the discriminative client model using authentic and imposter data of each client. Speaker recognition using SVM was reported in [12]. Support vector machine can be defined as the optimal hyper plane,

$$\langle w, x\rangle + b = 0, \quad (1)$$

that maximizes the distance of the separating hyper plane from the closest training data point called the support vectors. Here, w and b characterizes the direction and position in space, respectively and w is normal to the plane. For each direction, w, the hyperplane has the same distance from the nearest points from each class and the margin is twice this distance.

In support vector machine, the use of kernel function for non-linear separation can be executed when the linear boundary is inappropriate. In this case, the SVM maps the input vector onto a manifold embedded in a high dimensional feature space. In this study, polynomial kernel is employed. Three different types of single-sample systems have been developed and the information from zero, seven and eight verbal have been used as model in each single system for evaluation. The architecture of the single-sample systems is described as in Figure 2.

Multi-instance data fusion considers a combination of scores from several samples of different verbal that are extracted from the same modality. The overall architecture of the system is illustrated in Figure 3. Many studies revealed that integrating the scores of multiple samples can boost the performance of biometric systems.

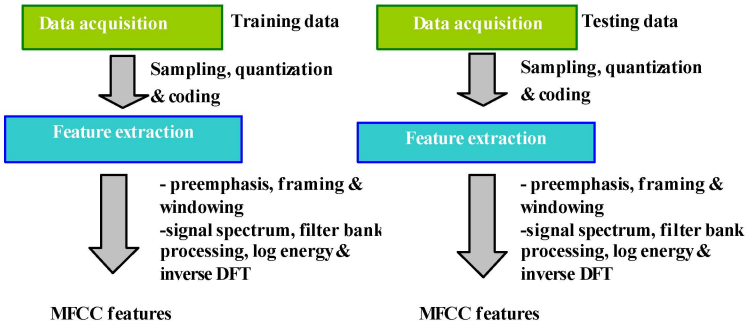

Fig. 1: Audio front end modules

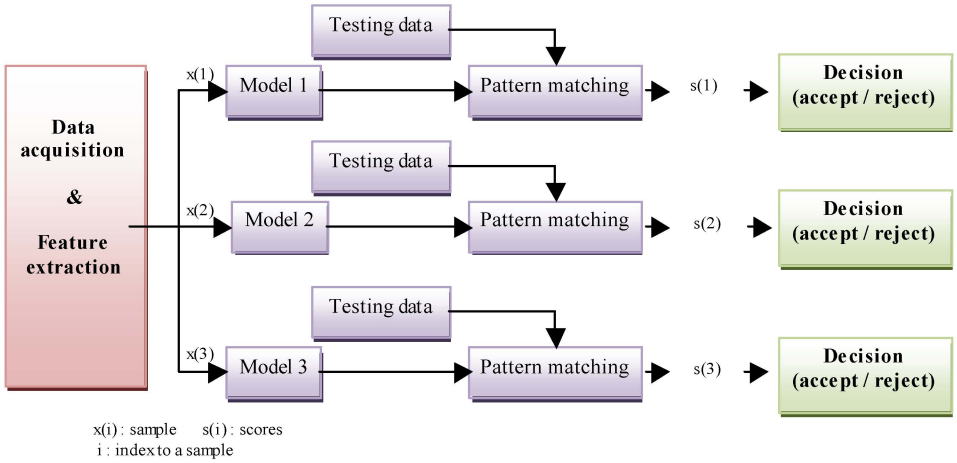

Fig. 2: Single-sample systems

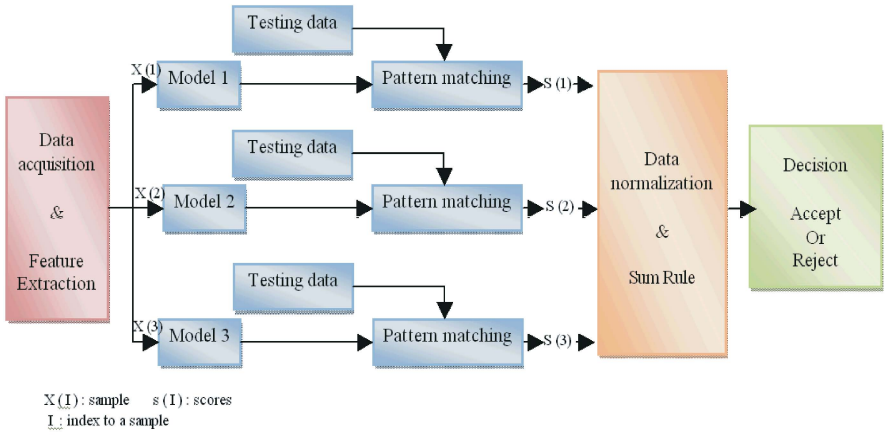

Fig. 3: Multi-instance systems

Kuncheva [13] investigated six operators for scores combining i.e. average, median, majority vote, maximum, minimum and Oracle. Among the six operators, performance using the average operator outperforms the other operators.

Implementing [14, 15] to this study, the scores for each verbal of all data are denoted as s_{ia}: $I = 1,..., k$. $a = 1,..., m$. denotes the ith match score output of each verbal type. k is the number of the match scores of each verbal and m is number of verbal type. The min-max normalized score, $\hat{s}_{ia}$ computed from each verbal score is given by

$$\hat{s}_{ia} = \frac{s_{ia} - \min_{i=1}^{n} s_{ia}}{\max_{i=1}^{n} s_{ia} - \min_{i=1}^{n} s_{ia}} \quad (2)$$

Where n is the number of the match scores available in the set. Let denote, $\hat{s}_{i1}$ $\hat{s}_{i2}$ and $\hat{s}_{i3}$ as verbal normalized, these normalized scores are then combined by using a Sum Rule scheme,

$$\hat{s}_{Fi} = \frac{\sum_{a=1}^{m} \hat{s}_{ia}}{m} \quad (3)$$

The scores of the single and multi-instance systems are then evaluated under Receiver Operating Curve (ROC) by for system performances.

RESULTS AND DISCUSSIONS

Figure 4 shows the performances of the single-sample system using data zero based on 20 training data, 10 training data, 6 training data and 3 training data. The increment of the numbers of training data gives a large improvement in the GAR performances. For example, at a FAR of 1%, the GAR of the 3, 6, 10 and 20 training data systems are 88%, 97%, 98% and 99 %, respectively. For other comparison, this study observes that the performance for 20, 10, 6 and 3 training data system achieves nearly 100% GAR at FAR of 5%, 10%, 20% and 60%, respectively. The system performances based on EER are shown in Table 1.

The performances of the single-sample system using data seven based on 20 training data, 10 training data, 6 training data and 3 training data are shown in Figure 5.The increase of the numbers of training data gives a good progress in the GAR performances. For instance, at a FAR of 1%, the GAR of the 3, 6, 10 and 20 training data systems are 73%, 83%, 86% and 88 %, respectively. Table 2 shows the system performances based on EER.

For the single-sample system using data eight, the system performances based on 20 training data, 10 training data, 6 training data and 3 training data are shown in Figure 6. The addition of the numbers of training data gives a good improvement in the GAR performances. In this case, at a FAR of 1%, the GAR of the 3, 6, 10 and 20 training data systems are 73%, 78%, 85% and 91 %, respectively. The system performances based on EER are shown in Table 3.

Table 1: EER performances for data zero at different numbers of training data

No. of training data	3	6	10	20
EER	4.3206	1.8168	1.5700	1.1524

Table 2: EER performances for data seven at different numbers of training data

No. of training data	3	6	10	20
EER	10.4148	9.7391	8.4535	7.9992

Table 3: EER performances for data eight at different numbers of training data

No. of training data	3	6	10	20
EER	9.6181	8.2320	6.6742	4.9916

Table 4: EER performances of single-instance systems and multi-instance system

Systems	Multi-instance	Single-sample zero	Single-sample seven	Single-sample eight
EER	2.0261	4.3206	10.4148	9.6181

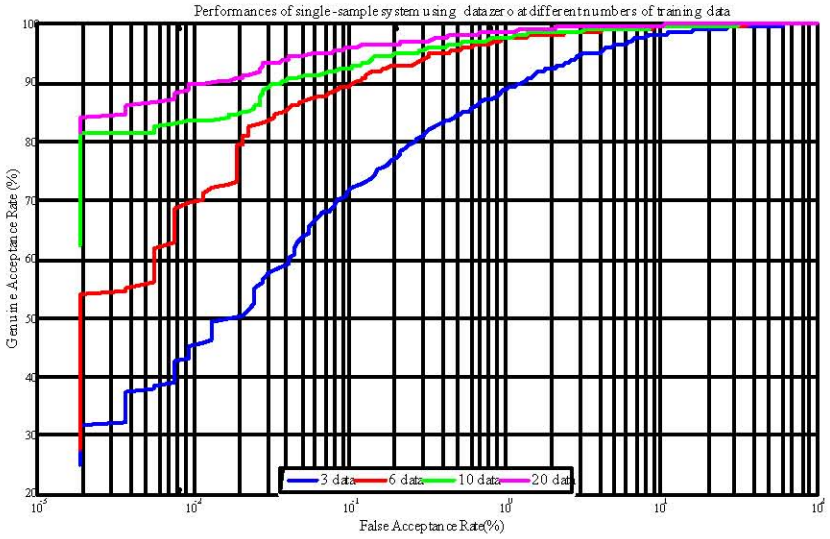

Fig. 4: Performance of single-sample system using data zero at different numbers of training data

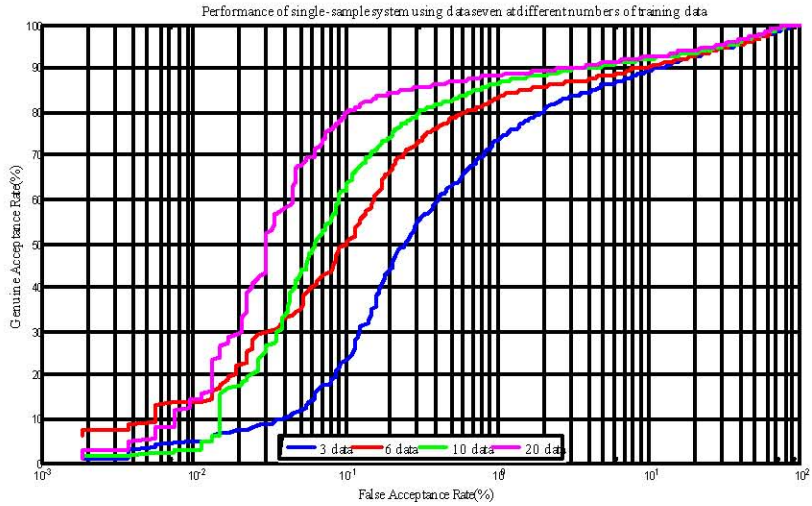

Fig. 5: Performance of single-sample system using data seven at different numbers of training data

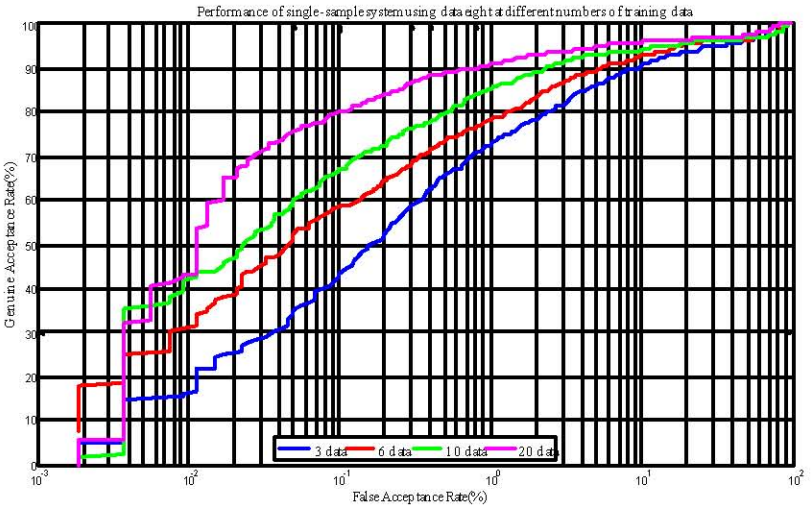

Fig. 6: Performance of single-sample system using data eight at different numbers of training data

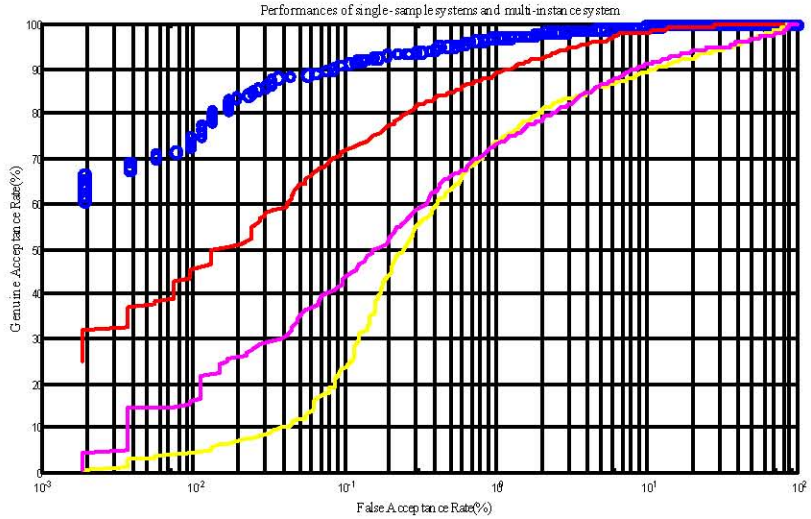

Fig. 7: Performance of single-instance systems and multi-instance system

Finally, Figure 7 compares the performances of all the single-sample systems with the multi-instance data fusion system based on 3 training data. At FAR of 1%, the GAR of the single-instance system for data zero, data seven and data eight are observed as 88%, 73% and 73 %, respectively. A good improvement is observed after implementing multi-instance data fusion i.e. the GAR of the multi-instance is observed as 97% at FAR 1%. For other comparison, this study observed the performance for multi-instance achieves nearly 100% GAR at FAR of 5% compared to 96% and 85% GAR for the single-sample systems. The system performances based on EER are shown in Table 4.

CONCLUSION

The performances of single-sample and multi-instance systems have been evaluated in this study. From the experimental results, the performance of multi-instance system outperforms the performances of all three single-sample systems. From this observation, implementing fusion technique is imperative in enhancing the performance of biometric system. Although, the choice of the verbal modeling has been done randomly in this study, the implication of the performances is still promising. Future research will be devoted to selection of the potential verbal as multi-instance samples to the multi-instance system.

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. Ross, A. and A.K. Jain, 2007. Fusion Techniques in multibiometric systems. In: R.I. Hammoud, B.R. Abidi and M.A. Abidi, Face Biometrics for Personal Identification, Berlin: Springer-Verlag Inc., pp: 185-212.
2. 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.
3. 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.
4. 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.
5. Teoh, A., S.A. Samad and A. Hussein, 2004. Nearest Neighbourhood Classifiers in a Bimodal Biometric Verification System Fusion Decision Scheme, Journal of Research and Practice in Information Technology, Australian Computer Society, 36(1): 47-62.
6. Ramli, D.A., S.A. Samad and A. Hussain, 2010. A Correlation Filter Based Biometric Speaker Authentication Systems, World Appl. Sci. J., 9(3): 259-267.
7. Keyhanipour, A.H., M. Piroozmand and K. Badie, 2009. A Neural Framework for Web Ranking Using Combination of Content and context Features, World Appl. Sci. J., 6(1): 6-15.
8. 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 Appl. Sci. J., 7(Special Issue of Computer & IT): 98-106.
9. Prabhakar, S. and A.K. Jain, 2000. Decision-level fusion in fingerprint verification. Pattern Recognition, 55(4): 861-874.
10. Jang, J., K.R. Park and Y. Lee, 2004. Multi-unit iris recognition system by image check algorithm. Proceeding of International Conference on Biometric Authentication (ICBA), pp: 450-457.
11. Furui, S., 2000. Digital speech processing, synthesis and recognition. USA: Marcel Dekker, Inc.
12. Wan, V. and W.M. Campbell, 2000. Support Vector Machines for speaker verification and identification. Proceeding of Neural Network for Signal Processing, 2: 775-784.
13. 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.
14. 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.
15. Ramli, D.A., S.A. Samad and A. Hussain, 2009. A Multibiometric Speaker Authentication System with SVM Audio Reliability Indicator. IAENG International J. Computer Sci. (IJCS-special issues), 36(4): 313-321.