Audio-Visual Based Multi-Sample Fusion to Enhance Correlation Filters Speaker Verification System

Dzati Athiar Ramli[1]
[1]School of Electrical & Electronic Engineering
USM Engineering Campus, University Sains Malaysia
14300, Nibong Tebal, Penang, Malaysia
dzati@eng.usm.my

Salina Abdul Samad[2], Aini Hussain[2]
[2]Faculty of Engineering
Universiti Kebangsaan Malaysia
43600, Bangi, Selangor, Malaysia

***Abstract*—In this study, we propose a novel approach for speaker verification system that uses a spectrogram image as features and Unconstrained Minimum Average Correlation Energy (UMACE) filters as classifiers. Since speech signal is a behavioral signal, the speech data has a tendency not to consistently reproduce due to the change of speaking rates, health, emotional conditions, temperature and humidity. In order to overcome this problem, a modification of UMACE filters architecture is proposed by executing a multi-sample fusion using speech and lipreading data. So as to evaluate the outstanding fusion scheme, five multi-sample fusion strategies, i.e. maximum, minimum, median, average and majority vote are first experimented using the speech signal data. Afterward, the performance of the audio-visual system using the enhanced UMACE filters is then tested. Here, lipreading data is combined to the audio samples pool and the outstanding fusion scheme that found in prior experiment is used as multi-sample fusion scheme. The Digit Database had been used for performance evaluation and the performance up to 99.64% is achieved by using the enhanced UMACE filters for the speech only system which is 6.89% improvement compared with the base line approach. Subsequently, the implementation of the audio-visual system is observed to be significant in order to broaden the PSR score interval between the authentic and imposter data as well as to further improve the performance of audio only system that offer toward a robust verification system.**

Keywords-multi-sample fusion, correlation filter,spectrographic image, lipreading, speaker verification

I. INTRODUCTION

Biometric speaker verification is generally a process of authenticating a speaker's claimed identity based on the speaker's voice [1]. The advantages of using speech signal information for biometric systems are that it is natural and easy to produce, requiring little custom hardware, has low computation requirement and is highly accurate (in clean noise-free conditions). But, since voice is categorized as a behavioral signal, the information has a tendency to be different with time and is not consistently reproduced due to the problems for instances the change of speaking rates, health, emotional conditions of speakers, temperature and humidity. Different microphones and channels as well as the limitation of the feature extractor and classifier also affect the accuracy of the system performance [2,3]. As a consequence, the execution of biometric systems has to properly distinguish the biometric features from one individual to another, and together, the system also needs to deal with the distortions in the features due the problem stated above.

So far, there are many advances have been explored in biometric systems so as to improve either the accuracy or the tolerance of the system to various inconsistent conditions. Fusion technique is one of the methods in literature that have been implemented in biometric systems so as to enhance the biometric systems performance. In general, the fusion method can be employed by combining several modalities (multi-modal fusion), combining several classifiers (multi-classifier fusion) and combining several samples of a single biometric modality (multi-sample fusion) [4]. Multi-modal fusion approach was described by Teoh et. al. in [5]. They proposed a combination of features from face modality and speech modality to improve the accuracy of biometric systems. Person identification based on visual and acoustic features has also been reported by Brunelli and Falavigna [6]. Combining the scores from different classifiers has been explained in [7,8]. The use of LVQ and MLP classifiers in footstep profile based person identification was reported by Suutala and Roning [7], while Kittler et.al. [8] utilized Neural Networks and HMM for the recognition of hand written digit. Finally, the implementation of fusion method using multi-sample derived from the same modality can be found in [4,9,10]. The papers showed that combining the scores of multiple samples can boost the biometric system greatly. Furthermore, Poh et al. [4] conclude that when two or more scores of a single modality biometric are averaged, noise that occurs due to classification can be reduced to the N (number of samples) factor.

This paper exploits the advantage of multi-modal fusion and multi sample fusion by considering several samples extracted from audio and lipreading modality as independent samples. By modifying the architecture of the original UMACE filter classifier, the samples are feed to the classifier as testing data. Although this technique employs many data samples but it does not impose any burden on users during data collection because a single and long sample of an utterance and a sequence of image from speakers can be simply separated into

a number of short samples [4,9,10]. Apart from that utilizing different modalities can assist the overall system to maintain a good performance because when one of the modality is corrupted, features from the other modality can still give a correct verification as described in [5,6].

In this study, we propose a novel approach by implementing the Unconstrained Minimum Average Correlation Energy (UMACE) filters for classification of speech signal. Here, we introduce a new feature i.e. a modified version of the speech signal in the form of its spectrographic image as features to the system since UMACE filters require 2-D image representation. So far, UMACE filters have been successfully applied for visual-based biometric recognition system i.e. face verification and fingerprint verification as described in [11] and [12], respectively. According to Savvides et al. [11] and Venkataramani et. al. [12], the special characteristics of UMACE filters are shift-invariance and ability to trade-off between discrimination and distortion tolerance. Other implementations of UMACE filters based on visual representation are person identification using lip motion sequence [13] and lower face verification [14].

The first objective of this study is to evaluate the performance of UMACE filter classifier and the proposed features i.e. spectrographic image for speaker verification system. The second objective is to determine the most appropriate fusion scheme for the enhancement of the original UMACE filter architecture so as to improve the performance of the base line system. Five multi-sample fusion strategies i.e. maximum, minimum, median, average and majority vote are experimented in order to evaluate the performance of the schemes. Finally, the performance of the enhanced UMACE filter using the audio and lipreading multi-sample data is then evaluated. The median and average operators are chosen as the fusion scheme for this task since they are found as the outstanding fusion scheme from the prior experiment.

II. MULTI-SAMPLE FUSION SCHEMES

A choice of an appropriate fusion method can further improve on the performance of the combination as stated in [15]. Combination of the individual outputs from the multi-sample scores can give a higher accuracy compared to the single output performances as reported in the study of multi-sample fusion strategies in [15]. The implementation of multi-sample fusion scheme on spectrographic and lipreading images is described as follows.

From [16], by expending the multi-sample theory to the lipreading application, assume that N streams of images, in our case spectrographic images and lipreading images, are extracted from M utterances / lipreading sequences $U = \{U_1,..., U_M\}$. We denote the spoken word / lipreading images corresponding to utterance U_m by

$$U^{(m)} = \left\{U_n^{(m)} \in \Re; n = 1,..., N_m\right\} \quad m = 1,..., M \tag{1}$$

where N_m is the number of spoken words / lipreading images in $U^{(m)}$ and n is the word / sequence index. In our experiment, the number of words employed is fixed to ten (zero to nine) while ten images from each lipreading sequence are employed. To simplify the notation, the M utterances / sequences contain the same number of spoken word / lipreading samples, i.e. $N = N_1 = N_2 = ... = N_m$.

From (1), assume the score for every sample from one utterance / lipreading sequence is denoted as $s_n; n = 1,..., N$. Let $s = \{s_1, s_2,..., s_N\}$ be a set (pool/ committee/ ensemble/ team) of scores from each utterance / sequence. The overall scores can be represented as

$$s(S_n; \Lambda) = \left\{s_n^{(1)},..., s_n^{(M)}; \Lambda\right\}, n = 1,..., N \tag{2}$$

containing the N spoken words / lipreading images from the M utterances / sequences.

Then, by considering each utterance / sequence, let's define $\hat{F} = f(s_1, s_2,..., s_N)$ as the fused estimate score. f is defined as the chosen fusion method. From [16], five fusion schemes are derived as describe below.

- For the maximum operator, the fused estimate score is decided by the maximum of $s = \{s_1, s_2,..., s_N\}$.

  $$\hat{F} = \max\{s_1, s_2,..., s_N\} \tag{3}$$

  The fused scores $\hat{F}$ are then compared against the decision threshold for the decision.

- For the minimum operator, the fused estimate score is decided by the minimum of $s = \{s_1, s_2,..., s_N\}$.

  $$\hat{F} = \min\{s_1, s_2,..., s_N\} \tag{4}$$

  The fused scores $\hat{F}$ are then compared against decision threshold for decision.

- For the median operator, the fused estimate score is decided by the median of $s = \{s_1, s_2,..., s_N\}$.

  $$\hat{F} = \text{med}\{s_1, s_2,..., s_N\} \tag{5}$$

  The fused scores $\hat{F}$ are then compared against decision threshold for decision.

- For the average operator, the fused estimate score is decided by the following equation:

  $$\hat{F} = \frac{1}{N}\sum_{n=1}^{N} s_n \tag{6}$$

  The fused scores $\hat{F}$ are then compared against decision threshold for decision.

- For the majority vote operator, the fused estimate score is decided by first assigning the individual scores,

  $$s_n \to 1 \text{ if } s_n \geq T \text{ or } s_n \to 0 \text{ if } s_n < T \tag{7}$$

  T is a threshold value. The decision is simply made by summing the binary values received. The class label

which is most represented among the N label outputs is chosen as a majority decision.

III. SPECTROGRAPHIC FEATURES EXTRACTION

Previously, human experts manually analyzed spectrogram in order to execute speaker recognition as describe in [17]. Spectrogram is a voiceprint image representing time-varying spectrum of a speech signal. The vertical axis (y) shows frequency while the horizontal axis (x) represents time. The pixel intensity or color represents the amount of energy (acoustic peaks) in frequency band y, at time x with red signifying the highest energies, then followed by orange, yellow, green, cyan, blue and magenta with gray areas having less energy and white areas below a threshold decibel level. It is simply an image formed by the magnitude of a short-time Fourier Transform [17,18]. Fig. 1 shows samples of spectrogram of the word 'zero' randomly taken from the database that we used in this study. From the figure, it is clear that the spectrogram image contains personal information in terms of the way the speaker utters the word such as speed and pitches that shown by the spectrum.

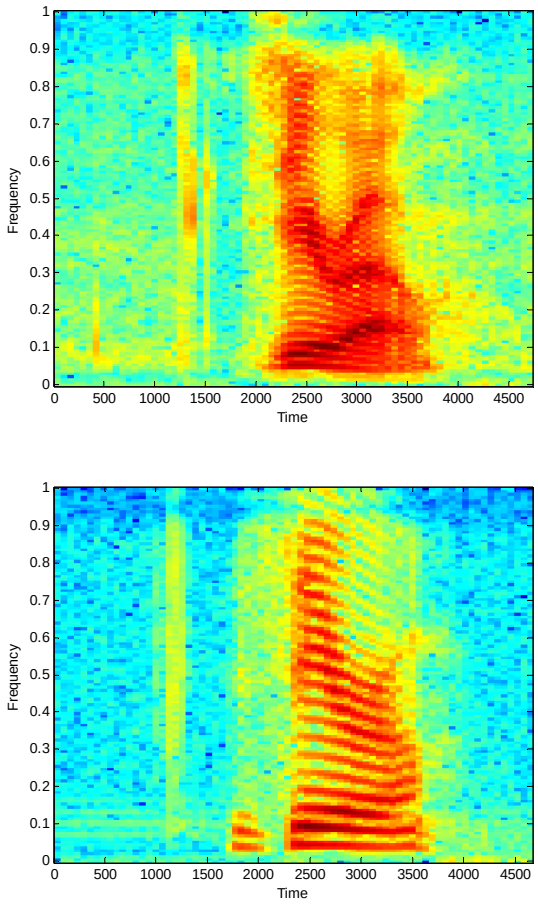

Figure 1. Example of the spectrogram images for word 'zero'

By comparing both images, although the spectrogram image holds inter-class variations for example the amount of high energies between individuals, it also comprises intra-class variations that affected by the low energies. Because of that, we propose a novel feature extraction technique in order to extract the optimum features from the spectrogram image so that it can be effectively classified by UMACE filter classifier. In general, the optimum features of the spectrogram, namely a spectrographic image can be derived by the following steps.

A. *Computation of the spectrogram*

The computation of the spectrogram is first executed the pre- emphasis task. By using a high-pass filter, the speech signal is filtered using the following equation:

$$x(t) = (s(t) - 0.95) * x(t-1) \quad (8)$$

$x(t)$ is the filtered signal, $s(t)$ is the input signal and t represents time. Followed by the framing and windowing task, a Hamming window with 20ms length and 50% overlapping is used on the signal. Fast Fourier Transform is then specified. A 256-point FFT is used and this value determines the frequencies at which the discrete-time Fourier transform is computed. Finally, the spectrum computation is executed. The logarithm of energy (acoustic peak) of each frequency bin is computed in this task.

B. *Excluding the low energies*

Excluding the low energies is a process to eliminate the small blobs in the image which impose the intra-class variations. In this step, the low energies of the acoustic peak are eliminated by setting an appropriate threshold to the FFT magnitudes during the computation of the spectrogram. Here, the FFT magnitudes which are above a threshold are maintained, otherwise they are set to be zero.

C. *Morphological image processing*

Morphological opening and closing process are then utilized so as to enhance the shape of the retained high energies. Morphological opening process is used to clear up the residue noisy spots in the image whereas morphological closing is the task to recover the original shape of the image caused by the morphological opening process.

Our spectrographic image database consists of 10 groups of spectrographic images (zero to nine) of 25 persons with 46 images per group of size 32x32 pixels, thus 11500 images in total. Fig. 2 shows the examples of spectrographic features from three different people randomly taken from the database.

Figure 2. Example of the spectrogram images for word 'zero'

IV. LIPREADING FEATURES EXTRACTION

Lipreading features extraction consist of two distinct steps i.e. face detection and lip localization. We use the methods that have been implemented in [19,20], in order to locate the lips on a face. In the first step, a color-based technique and template matching algorithm are employed to segment human skin regions from non-skin color. The Gaussian model $N(\mu, C)$ with mean vector $\mu = E[x]$ and covariance matrix $C = E[(x-\mu)(x-\mu)^T]$ which are obtained from different skin color images has been used to determine the skin likelihood for any pixel of an image.

The skin likelihood is computed as $P(r,b) = \exp[-0.5(x-m)^T C^{-1}(x-m)]$ where $x = (r,b)^T$ i.e. chromatic pair value of red and blue. The skin-likelihood image is transformed to the skin-segmented image (binary image) as depicted in Fig. 3 and Fig. 4, respectively.

Figure 3. Skin-likelihood image

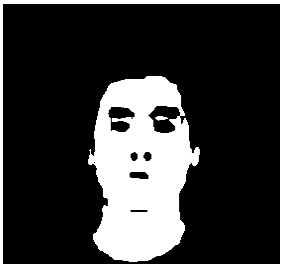

Figure 4. Skin-segmented image

For the lip localization task, hue/saturation color thresholding has been used in order to differentiate the lip area from the face as explained in [19,20]. Lip detection by using hue/saturation color is much easier due to its robustness under wide range of lip colors and varying illumination condition according to Matthews et al. [21]. The example of hue-saturation image is shown in Fig. 5. From the hue-saturation image, a binary image is then computed by setting the threshold values, $H_0 = 0.04$ and $S_0 = 0.1$ as shown in Fig. 6. By applying morphological image processing, the largest blob is determined as a lip region. The lip regions of 64 x 64 pixels are then extracted for evaluation.

Figure 5. Hue-saturation image

Figure 6. Binary image for lip region localization

Our lipreading image database consists of 41 sequences of lipreading images from 25 persons with 20 images per sequence of size 64x64 pixels, thus 20500 images in total. Fig. 7 shows the examples of spectrographic features from three different people from the database.

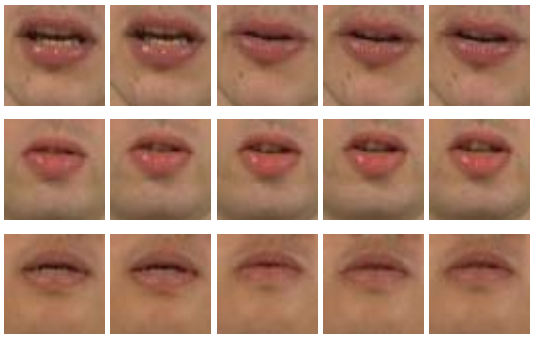

Figure 7. Example of the lipreading images from lipreading sequence

V. VERIFICATION USING CORRELATION FILTERS

One of the advanced correlation filter, namely the Unconstrained Minimum Average Correlation Energy (UMACE) filters are implemented as classifier to the system in this study. These advanced correlation filters are evolved from Matched Filters which are optimal for detecting a known reference image in the presence of additive white Gaussian noise [11,12]. The particular characteristic of these advanced correlation filters are tolerant in the occurrence of distortions such as illumination changes and facial expression as described in Savvides et al. [11]. These special advantages offer an attractive technique for our application in handling intra-class variations in spectrographic and lipreading images.

The filters are synthesized in the Fourier domain using closed form equations. Several training images are used to synthesize a filter template. The designed filter is then used for cross-correlating the test image in order to determine whether the test image is from the authentic class or imposter class. In this process, the filter optimizes a criterion to produce a desired correlation output plane by minimizing the average correlation energy and at the same time maximizing the correlation output in the origin. The resulting correlation plane produce a sharp peak in the origin and the values at everywhere else are close to zero when the test image belongs to the same class of the designed filter [11,12]. Fig. 8 shows the correlation outputs when using a UMACE filter to determine the test image from the authentic class (left) and imposter class (right).

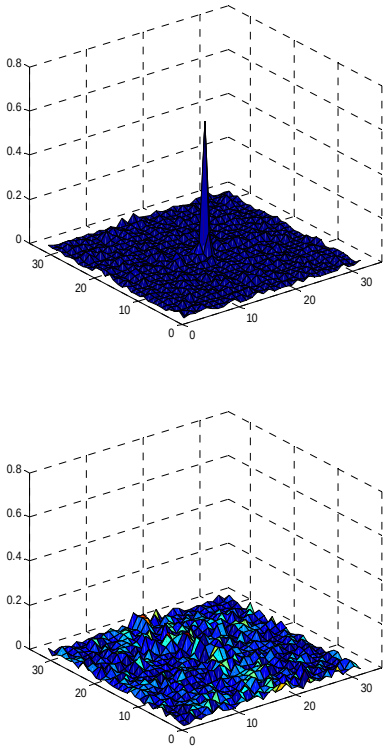

Figure 8. Examples of the correlation plane for the test image from the authentic class (left) and imposter class (right).

The optimization of UMACE filters equation can be summarized as follows.

$$U_{mace} = D^{-1}m \tag{9}$$

D is a diagonal matrix with the average power spectrum of the training images placed along the diagonal elements while ***m*** is a column vector containing the mean of the Fourier transforms of the training images. Peak-to-Sidelobe ratio (PSR) metric is used to measure the sharpness of the peak. The PSR is given by

$$PSR = \frac{peak - mean}{\sigma} \tag{10}$$

Here, the peak is the largest value of the test image yield from the correlation output. Mean and standard deviation are calculated from the 20x20 sidelobe region by excluding a 5x5 central mask [11].

VI. SPEAKER VERIFICATION SYSTEM USING ENHANCED UMACE FILTERS

Audio-Visual Digit Database (2001) developed by Sanderson is used for the purpose of this study (2001) [22]. The database consists of video and the corresponding audio of people reciting digits zero to nine. The audio provided is a monophonic, 16 bit, 32 kHz, WAV format.

The enhanced UMACE filter for spectrographic verification is designed as in Fig. 9. Since, we consider 10 groups of digits (zero to nine) as samples; ten filters are designed for each person in the database. Our spectrographic image database consists of 46 images per group of size 32x32 pixels of 25 persons, thus 11500 images in total. For each filter, we used 6 training images for the synthesis of the UMACE filter. Then, 40 images are used for the testing process. These six training images were chosen based on the largest variations among the images. In the testing stage, we performed cross correlations of each corresponding word with 40 authentic images and another 40x24=960 imposter images from the other 24 persons.

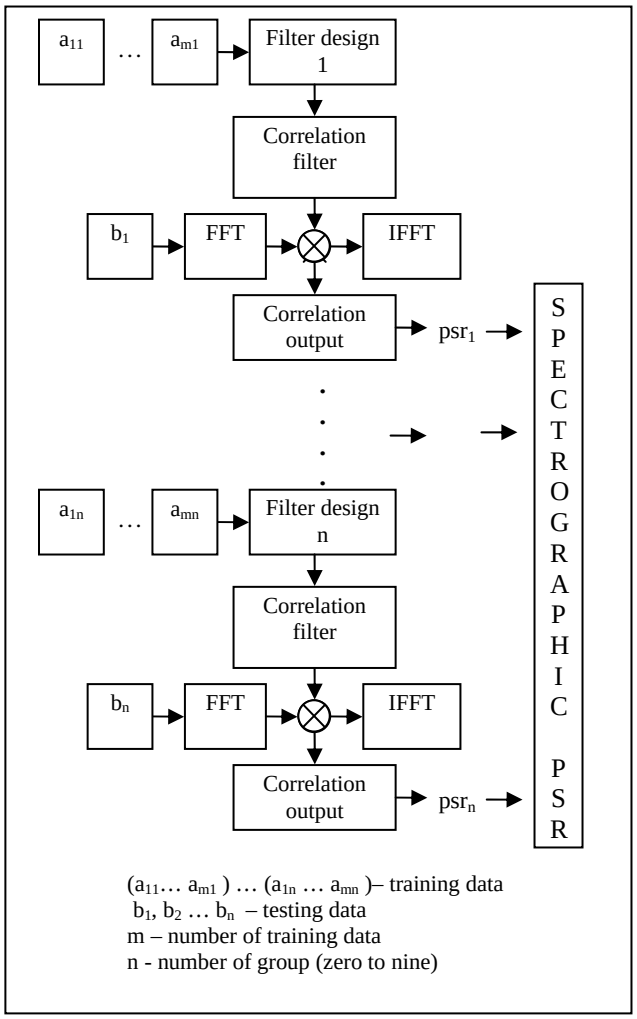

Figure 9. Verification for spectrographic images using enhanced correlation filters.

Our lipreading database consists of 20x25x41 =20500 localized lip images. 41 sequences of frames from 25 people have been extracted while each sequence consists of 20 images. Training images that consist of 6 images which to be synthesized for each person UMACE filters' are chosen from one of the 41 sequences. In our case, we have 25 filters which represent each person in the database. During the testing stage, we performed cross correlations of each person by using their corresponding filter with 200 authentic images (40 authentic sequences) and another 4800 imposter images (40x24=960 imposter sequences) from the other 24 persons. Fig. 10 illustrates the details of the lipreading verification process using the enhanced UMACE filters architecture.

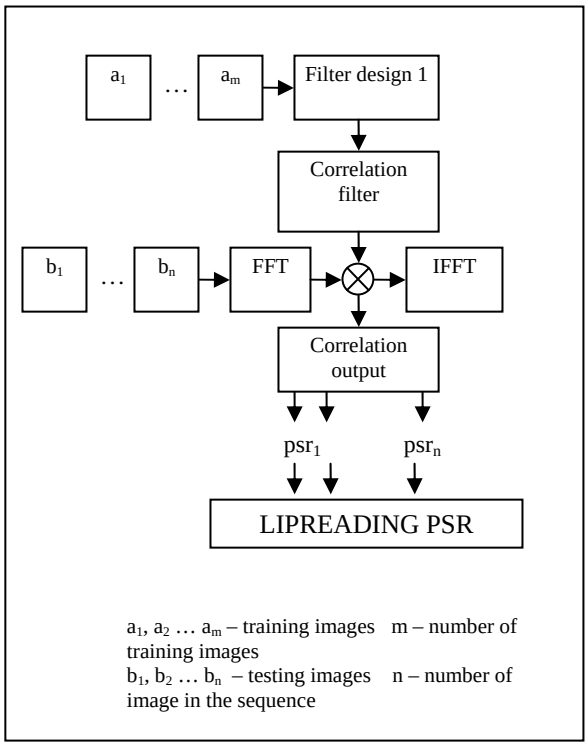

Figure 10. Verification for lipreading images using enhanced correlation filters.

For the audio based system, psr values from ten spectrographic samples (zero to nine) are used for fusion process meanwhile for audio-visual based system, psr values from five spectrographic samples (zero to four) and five lipreading samples are employed. The process of audio and audio-visual based system is illustrated in Fig. 11 and Fig. 12, respectively.

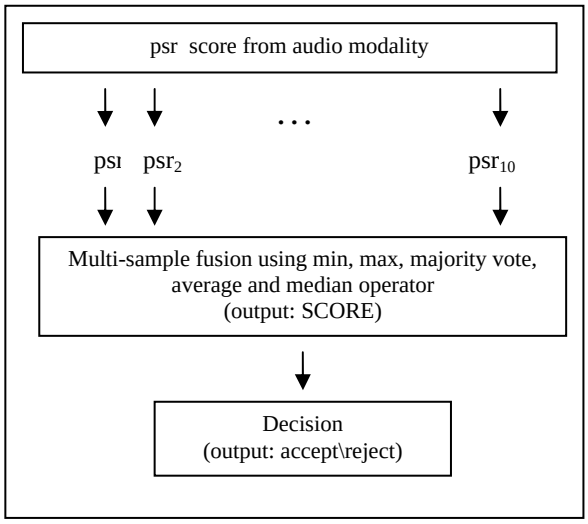

Figure 11. Decision process for audio based system.

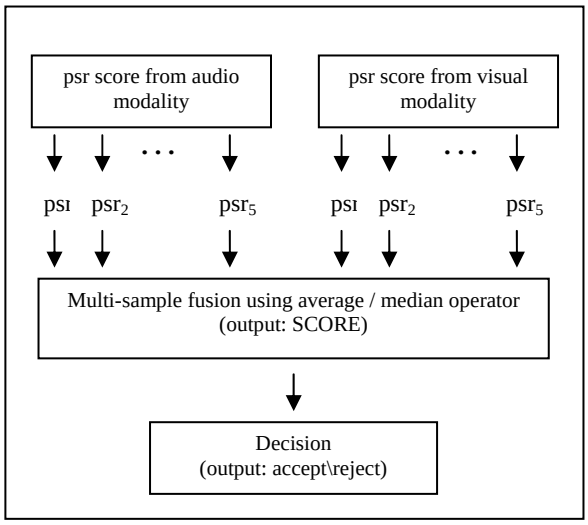

Figure 12. Decision process for audio-visual based system.

VII. RESULTS AND DISCUSSIONS

By cross-correlating all the spoken word and lipreading images with the respective filters in the database, the performance of each person's UMACE filters can be assessed Then, their corresponding psr and fusion score values (SCORE) for each scheme are computed and recorded. The fusion scores are then compared to their corresponding stored threshold values for the decision. In this study, the performance evaluations are based on false acceptance rate (FAR) and false rejection rate (FRR) which calculated as defined below.

$$\text{FAR} = \frac{\text{Number of imposters}\,(\text{SCORE} > S_0)}{\text{Total imposters}} \qquad (11)$$

$$\text{FRR} = \frac{\text{Number of authentic}\,(\text{SCORE} < S_0)}{\text{Total authentics}} \qquad (12)$$

Then, overall performance is calculated by combining these two errors into total success rate (TSR) where

$$\text{TSR} = 100\% - \left(\frac{\text{FAR} + \text{FRR}}{\text{Total number of accesses}}\right)100\% \qquad (13)$$

Table 1 below compares the system performance of UMACE filter and modified UMACE filter via median operator. An improvement by 6.89% is achieved by implementing the fusion approach.

TABLE I. TSR PERCENTAGES USING UMACE FILTER AND MODIFIED UMACE FILTER (MEDIAN OPERATOR)

features	UMACE	modified UMACE
spectrographic	92.75	99.64

Table 2 describes the overall system performance by executing five multi-sample fusion schemes by calculating the FAR, FRR and TSR percentages. From the experiment, we found that median and average operators are the most outstanding operators based on lower FAR and FRR percentages.

TABLE II. SYSTEM PERFORMANCE PERCENTAGES BASED ON FIVE MULTI-SAMPLE FUSION SCHEMES

scheme	FAR %	FRR %	TSR %
maximum	0.93	11.8	98.63
minimum	1.3	25.7	97.72
median	0.25	2.9	99.64
average	0.2	5.1	99.6
majority vote	0.99	6.6	98.78

FAR and FRR percentages of audio and audio-visual system using median and average operator are described in Table 3. The results show that the system performance increases after combining visual psr values to the audio sample pool. We also observe that a large gap of separation between the maximum SCORE values for the imposters and the minimum SCORE values for the authentic after implementing this technique. This is due to UMACE filters have been proven to perform well in visual implementation either using face or lip representation as found in [11],[12],[13] and [14]. In this experiment, the high psr values from authentic scores give advantage to the system to boost the system performance after the fusion process takes place.

TABLE III. ERROR PERCENTAGES BASED ON AUDIO AND AUDIO-VISUAL SYSTEM USING MEDIAN AND AVERAGE OPERATOR

	median		average	
	FAR	FRR	FAR	FAR
audio	0.25	2.9	0.2	5.1
audio-visual	0.1	0.16	0.09	0.18

Fig. 13 and Fig. 14 compare the psr performance of audio and audio-visual system using median operator, respectively. The psr performance using average operator is illustrated in Fig. 15 and Fig. 16. Larger margins of separation between the authentic and imposter scores are observed for the audio-visual based system compared to the audio based system.

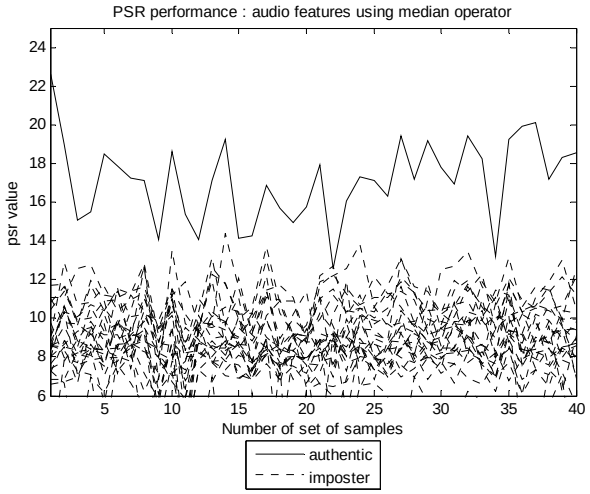

Figure 13. PSR performance for audio system using median operator.

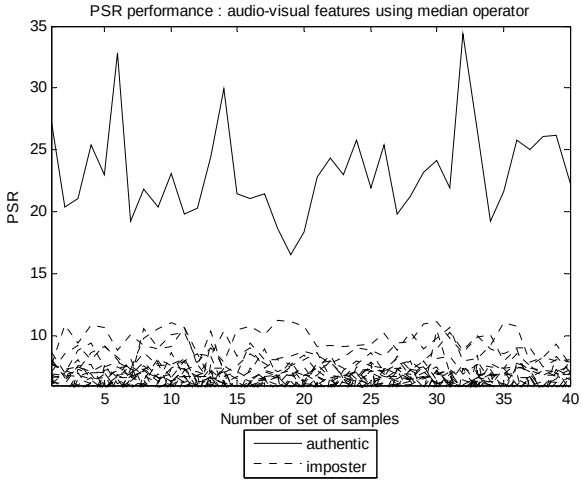

Figure 14. PSR performance for audio-visual system using median operator.

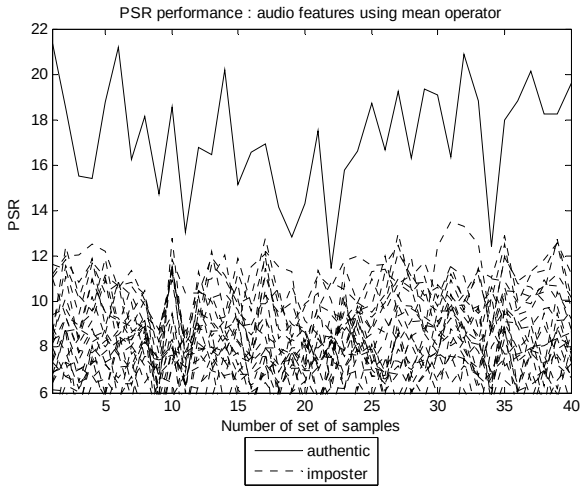

Figure 15. PSR performance for audio system using mean operator.

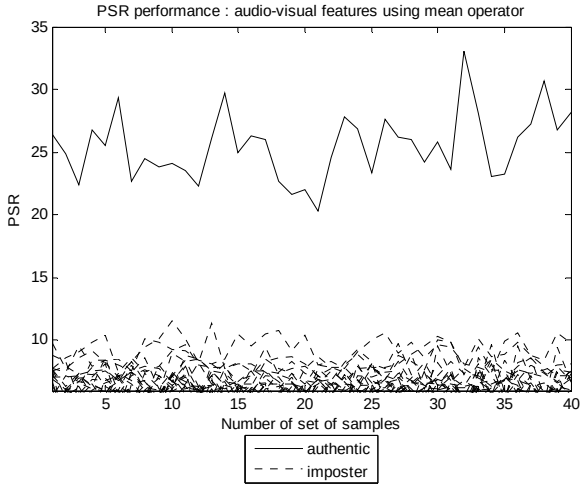

Figure 16. PSR performance for audio-visual system using mean operator.

VIII. CONCLUSIONS

A promising result from the baseline experiment shows that our proposed spectrographic features and UMACE filters classifier can be implemented as an alternative technique to perform biometric speaker verification system. Subsequently, the enhanced version of UMACE filters classifier using multi-sample fusion can further improve the system performance because by executing multi sample fusion, the error due to the variation of data can be reduced. Apart from that the finding also shows that the correct choice of operator is important in implementing the multi-sample fusion system. This study also reveals that the audio-visual system is significant to confirm a wide margin of separation between the authentic and imposter scores so that the system can minimize the FAR and FRR error percentages. This approach offers an advantage when the system is implemented in more adverse conditions. In such situation, when one of the modality is degraded, the other modality can assist to maintain the system performance.

ACKNOWLEDGMENT

This research is supported by the following research grants: Fundamental Research Grant Scheme, Malaysian Ministry of Higher Education, FRGS UKM-KK-02-FRGS0036-2006, Science Fund, Malaysian Ministry of Science, Technology and Innovation, 01-01-02-SF0374 and Incentive Grant Universiti Sains Malaysia.

REFERENCES

[1] J.P. Campbell, "Speaker Recognition: A Tutorial", Proceeding of the IEEE, Vol. 85, pp. 1437-1462, 1997.

[2] A. Rosenberg, "Automatic speaker verification: A review", Proceeding of IEEE, Vol. 64(4), pp. 475-487, 1976.

[3] D.A. Reynolds, "An overview of Automatic Speaker Recognition Technology", Proceeding of IEEE on Acoustics Speech and Signal Processing, Vol. 4, pp. 4072-4075, 2002.

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

[5] A. Teoh, S.A. Samad, A. Hussein, "Nearest Neighbourhood Classifiers in a Bimodal Biometric Verification System Fusion Decision Scheme", Journal of Research and Practise in Information Technology, Vol. 36(1), pp. 47-62, 2004.

[6] R. Brunelli, D. Falavigna, "Personal Identification using Multiple Cue", IEEE Transaction on Pattern Analysis and Machine Intelligence, Vol. 17(10), pp. 955–966, 1995.

[7] J. Suutala, J. Roning, "Combining Classifier with Different Footstep Feature Sets and Multiple Samples for Person Identification", Proceeding of International Conference on Acoustics, Speech and Signal Processing, pp. 357-360, 2005.

[8] J. Kittler, M. Hatef, R.P.W. Duin, J. Matas, "On combining classifiers", Proceeding of the IEEE Transaction On Pattern Analysis and Machine Intelligence, Vol. 20(3), pp. 226-239, 1998.

[9] M.C. Cheung, M.W. Mak, S.Y. Kung, "Multi-Sample Data-Dependent Fusion of Sorted Score Sequences for Biometric verification", Proceeding of the IEEE Conference on Acoustics Speech and Signal Processing (ICASSP04), pp. 229-232, 2004.

[10] M.C. Cheung, K.K. Yiu, M.W. Mak, S.Y. Kung, "Multi-Sample Fusion with Contrained Feature Transformation for Robust Speaker Verification", Proceeding of the IEEE Conference on Acoustics Speech and Signal Processing (ICASSP04), pp. 1813-1816, 2004.

[11] M. Savvides, B.V.K. Vijaya Kumar, P. Khosla, "Face Verification using Correlation Filters", Proceeding of Third IEEE Automatic Identification Advanced Technologies, pp. 56-61, 2002.

[12] K. Venkataramani, B.V.K. Vijaya Kumar, "Fingerprint Verification using Correlation Filters System", pp. 886-894, 2003.

[13] S.A. Samad, D.A. Ramli, A. Hussain, "Person Identification using Lip Motion Sequence", in Apolloni, B. et al. (eds), Springer-Verlag Berlin Heidelberg, LNAI, Vol. 4692, Part 1, pp. 839-846, 2007.

[14] S.A. Samad, D.A. Ramli, A. Hussain, "Lower Face Verification Centered on Lips using Correlation Filters", Information Technology Journal, Vol. 6(8), pp. 1146-1151, 2007.

[15] L.I. Kuncheva, "A theoretical Study on Six Classifier Fusion Strategies", Proceeding of the IEEE Transaction On Pattern Analysis and Machine Intelligence, pp. 348-353, 2001.

[16] Salina Abdul Samad, Dzati Athiar Ramli, Aini Hussain. Comparative Study on Several Multi-Sample Fusion Schemes for Speaker Verification System. WSEAS International Conference on Circuit, System, Electronics, Control & Signal Processing (CSECS 2007), ps. 335-340, 2007.

[17] R.L. Klevents, R.D. Rodman, Voice Recognition, Artech House, London, INC, 1997.

[18] http://cslu.cse.ogi.edu/tutordemo/spectrogramReading/spectrogram.html.

[19] G. Chetty, M. Wagner, "Liveness Verification in Audio-Video Speaker Authentication", Proceeding of International Conference on Spoken Language Processing ICSLP 04, pp. 2509-2512, 2004.

[20] G. Chetty, M. Wagner, "Automated Lip Feature Extraction for Liveness Verification", Proceeding of Image and Vision Computing, pp. 17-22, 2004.

[21] I. Matthews, J. Cootes, J. Bangham, S. Cox, R. Harvey, "Extraction of Visual Features for Lipreading", Proceeding of the IEEE Transaction. on Pattern Analysis and Machine Intelligence, Vol. 24(2), pp. 198–213, 2002.

[22] C. Sanderson, K.K. Paliwal, "Noise Compensation in a Multi-Modal Verification System", Proceeding of International Conference on Acoustics, Speech and Signal Processing, pp. 157-160, 2001.

AUTHORS PROFILE

Dzati Athiar Ramli received the B.App.Sc. and M.Sc. in applied mathematics from Universiti Sains Malaysia and Ph.D in software engineering from Department of Electrical, Electronic & Systems Engineering, Universiti Kebangsaan Malaysia. She is a lecturer at the School of Electrical & Electronic Engineering, Universiti Sains Malaysia Her research interests include speech and speaker recognition algorithm and software, multimodal biometric system, signal & image processing and engineering mathematics.

Salina Abdul Samad received the BSEE and Ph.D. in electrical engineering from University of Tennessee and University of Nottingham, respectively. She is a professor at the Department of Electrical, Electronic & Systems Engineering, Universiti Kebangsaan Malaysia. Her research interests include digital signal processing, filter design and multimodal biometric system.

Aini Hussain received the BSEE, M.Sc. and Ph.D. in electrical engineering from Louisiana State University, UMIST and Universiti Kebangsaan Malaysia, respectively. She is a professor at the Department of Electrical, Electronic & Systems Engineering, Universiti Kebangsaan Malaysia. Her research interests include intelligent signal processing, pattern recognition and system modeling.