A Comparative Study on Multi-Sample Fusion Schemes to Enhance Spectrographic Speaker Verification

SALINA ABDUL SAMAD DZATI ATHIAR RAMLI AINI HUSSAIN
Department of Electrical, Electronic and Systems Engineering, Faculty of Engineering
University Kebangsaan Malaysia
43600 Bangi Selangor
MALAYSIA

Abstract: - In this paper, a comparative study using multi-sample fusion schemes is evaluated in order to enhance the performance of a speaker verification system. Five multi-sample fusion strategies, i.e. the maximum, minimum, median, average and majority vote are employed for this purpose. In this study, we propose a novel approach by using the spectrogram image of the spoken word as features and the Unconstrained Minimum Average Correlation Energy (UMACE) filters are utilized as classifiers to the system. The Digit Database is used for performance evaluation and results show that the median operator outperforms the other fusion schemes yielding an overall performance of 99.64%.

Key-Words: - multi-sample fusion, UMACE filter, spectrographic feature, speaker verification

1 Introduction

Biometric speaker verification is generally a task of accepting or rejecting a speaker's claimed identity based on the speaker's voice [1]. A major problem with biometric systems that utilize the information contained in speech signals is that the information has a tendency to vary with time and is not consistently reproduced. This is owing to the differences in speaking rates, health and emotional conditions of speakers. In addition, perfect verification may not be achieved due to different microphones and channels used as well as the limitation of the feature extractor and classifier [2][3]. Thus, the implementation of biometric systems have to appropriately discriminate the biometric features from one individual to another, and at the same time, the system also needs to deal with the distortions in the features due the problem stated above. One of the solutions that can be used to overcome these limitations is by executing fusion techniques in the system [4].

So far, there are many fusion techniques in literature that have been implemented in biometric systems for the purpose of enhancing the system performance. Generally, the fusion methods can be categorized into 3 major groups, i.e., multi-modal fusion, multi-classifier fusion and multi-sample fusion [4]. A study in multi-modal fusion approach was described by Teoh et. al. in [5]. They proposed a combination of features of face modality and speech modality so as to improve the accuracy of biometric systems. Person identification based on visual and acoustic features has also been reported by Brunelli and Falavigna in [6]. Their findings conclude that when one of the modalities is somehow degraded, the other modality can assist the system to maintain a good performance.

The second group of fusion techniques that combine the scores from different classifiers has been described in [7] and [8]. Suutala and Roning in [7] used LVQ and MLP as classifiers for footstep profile based person identification whereas in [8], Kittler et.al. utilized Neural Networks and HMM for the hand written digit recognition task. In both study, multi classifiers improved the system reliably.

Finally, for the multi-sample fusion approach, a combination of scores from several features of single modality is computed. The implementation of multi-sample fusion approach can be found in [4], [9] and [10]. The papers showed that combining the scores of multiple samples can boost the biometric system greatly.

This paper concentrates on the multi-sample fusion schemes by considering several samples extracted from the same modality as independent samples. There are three benefits of implementing multi-sample fusion. First, in term of features, a single and long sample of an utterance from speakers can be simply separated into a number of short samples. Although this technique employs many data samples but it does not impose any burden on users during data collection. Second,

similar features extraction process on every sample and utilizing only one classifier do not burden the system compared to the multi-modal system and multi-classifier system. Finally, the cost of executing multi sample system is lower because only one sensor is involved [10].

The objectives of this study are to compare the performance of multi-sample fusion system to single-sample system as well as to determine the most appropriate fusion scheme for our system. In this study, we propose a novel approach by implementing multi-sample fusion using the images of the spoken words from the spectrogram as features to a speaker verification system. This study was inspired by the publications of Kittler et al. [8] and Kuncheva [11] where classifier fusion strategies are experimentally compared. Five multi-sample fusion strategies i.e. maximum, minimum, median, average and majority vote are employed in order to evaluate the performance of the schemes.

The Unconstrained Minimum Average Correlation Energy (UMACE) filter is then executed as classifier to the system. UMACE filter has been successfully applied in visual-based biometric recognition system as well as in speech signal authentication. Person identification based on lip information using UMACE filter can be found in Samad et al. [12]. The performance of lower face verification using UMACE filter is also evaluated by Samad et al. in [13]. A study on voice-print analysis by using UMACE filter for single sample approach is reported in Ramli et al. [14].

2 Multi-Sample Fusion Schemes

A study on multi-sample fusion strategies can be found in [11]. By combining the individual outputs from the multi-sample scores, we aim at a higher accuracy than that of the best score. As stated in [11], a choice of an appropriate fusion method can further improve on the performance of the combination.

Assume that N streams of words (in our case spectrographic images) are extracted from M utterances $U = \{U_1,...,U_M\}$. We denote the spoken word sequence 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 in $U^{(m)}$ and n is the word index. In our experiment, the number of words employed is fixed to ten (zero to nine). To simplify the notation, the M utterances contain the same number samples, i.e. $N = N_1 = N_2 = ... = N_m$.

From (1), assume the score for every sample from one utterence 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. 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 from the M utterances.

Then, by considering each utterance, lets define $\hat{F} = f(s_1, s_2,..., s_N)$ as the fused estimate score. f is defined as the chosen fusion method. Subsequently, five fusion schemes are derived as follows.

2.1 Maximum operator

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.

2.2 Minimum operator

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.

2.3 Median operator

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.

2.4 Average operator

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.

2.5 Majority vote operator

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

scores,

$$s_n \to 1 \text{ if } s_n \geq T \text{ or } s_n \to 0 \text{ if } s_n < T \quad (7)$$

where 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.

3 Features Extraction

In the past, human experts manually interpret voiceprint for semiautomatic speaker recognition [15]. A spectrogram is used to represent the voiceprint is an image representing the time-varying spectrum of a signal. The vertical axis (y) shows frequency, the horizontal axis (x) represents time and the pixel intensity or color represents the amount of energy (acoustic peaks) in frequency band y, at time x [16]. Fig.1 and Fig.2 show samples of spectrogram of the word 'zero' from person 3 and person 4 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 pitch that is showed by the spectrum.

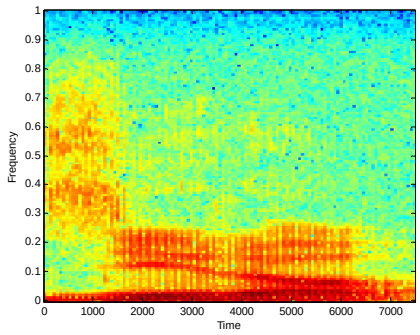

Fig.1 Example of the spectrogram image from person 3 for word 'zero'.

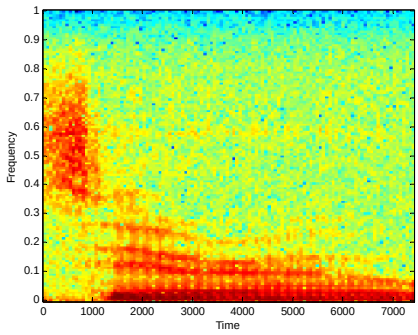

Fig.2 Example of the spectrogram image from person 4 for word 'zero'.

By comparing both figures, although the spectrogram image holds inter-class variations, it also comprises intra-class variations. In order to be successfully classified by UMACE filter, we propose a novel feature extraction technique.

In general, our features i.e. spectrographic image can be derived by the following steps.

3.1 Computation of the spectrogram

The computation of the spectrogram is described below. The input to the algorithm is a speech signal.

1. 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.

2. Framing and windowing task. A Hamming window with 20ms length and 50% overlapping is used on the signal.
3. Specification of FFT length. A 256-point FFT is used and this value determines the frequencies at which the discrete-time Fourier transform is computed.
4. The logarithm of energy (acoustic peak) of each frequency bin is then computed.

3.2 Retaining the high energies

After a spectrogram image is obtained, we aim to eliminate the small blobs in the image which impose the intra-class variations. This aim can be achieved by retaining the high energies of the acoustic peak by setting an appropriate threshold. Here, the FFT magnitudes which are above a threshold are maintained, otherwise they are set to be zero. Fig.3 shows the image with retained high energies.

Fig.3 Image from person 3 (left) and person 4 (right) after retaining the high energies

3.3 Morphological opening and closing

Next, morphological opening and closing process are utilized. 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. Fig.4 shows the final image to be used in the classification process.

Fig.4 Image from person 3 (left) and person 4 (right) after morphological opening and closing

4 Classification by UMACE Filters

The optimization of Unconstrained Minimum Average Correlation Energy (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.

UMACE filters which are evolved from Matched Filter 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 [17][18]. Fig. 5 shows the correlation outputs when using a UMACE filter to determine the test image from the authentic class (left) and imposter class (right). The verification process using UMACE filter is summarized in Fig. 6.

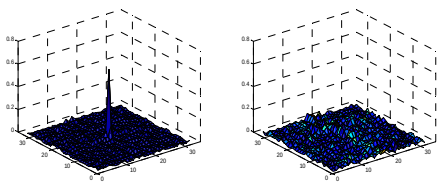

Fig.5 Examples of the correlation plane for the test image from the authentic class (left) and imposter class (right).

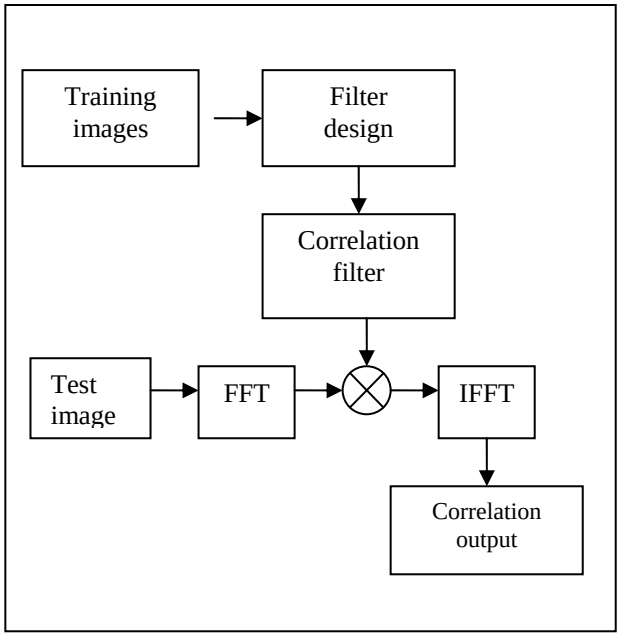

Fig.6 Verification process using UMACE filter.

According to Savvides et al. [17], these advanced correlation filters are tolerant in the presence of distortions such as illumination changes and facial expression. This special advantage offer an attractive technique for our application in handling intra-class variations of spectrographic 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 [17].

5 Speaker Verification System

Audio-Visual Digit Database (2001) developed by Sanderson is used for the purpose of this study (2001) [19]. 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.

In our experiments, we use 250 filters which represent each word for the 25 persons. 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. 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. The verification architecture of the system is shown in Figure 7.

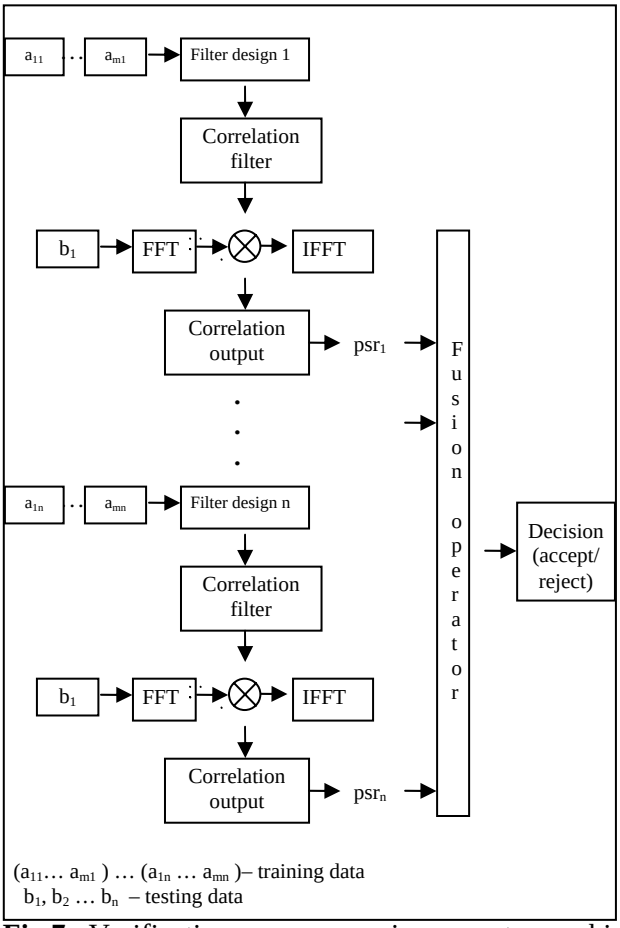

Fig.7 Verification process using spectrographic images.

6 Results and discussion

In the experiments, the performance of each person's UMACE filter are assessed by cross-correlating all the spoken word images in the database. Then, their corresponding PSR and fusion score values for each scheme are computed and recorded. The fusion scores are then compared to their corresponding stored threshold values for the decision. False acceptance rate (FAR) and false rejection rate (FRR) are then 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 single-sample approach and multi-sample fusion approach via median operator. An improvement by 6.89% is achieved by implementing the fusion approach.

Table 1. TSR percentages of single sample and multi sample scheme (median operator)

features	single sample	multi sample
spectrographic	92.75	99.64

Table 2 describes the overall system performance by executing five multi-sample fusion schemes. The performance of the system is based on FAR, FRR and TSR percentages.

Table 2. 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

7 Conclusion

Five multi-sample fusion schemes i.e. the maximum, minimum, median, average and majority vote, together with the single-sample approach are compared experimentally in this study. The findings showed that the multi-sample fusion approach is always superior compared to the single-sample approach, and median operator is found to be the best scheme to be implemented in our speaker verification system. An outstanding performance

especially in terms of FRR percentage of 2.9% for the median operator compared to other operators concludes that the correct choice of operator is important in implementing the multi-sample fusion system. Apart from that, this study also showed that the used of our novel approach that employs spectrographic images as features together with UMACE filter can be a promising alternative technique in a biometric speaker verification system. Our technique offers a robust feature extraction and classification process to treat the intra-class variations of speech signals.

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

References:

[1] J.P. Campbell, Speaker Recognition: A Tutorial, *in Proc. of the IEEE*, 1997, vol. 85, pp. 1437-1462.
[2] A. Rosenberg, Automatic speaker verification: A review, *in Proc. of IEEE*, 1976, vol. 64, no. 4, pp.475-487.
[3] D.A. Reynolds, An overview of Automatic Speaker Recognition Technology, *in Proc. of IEEE on Acoustics Speech and Signal Processing*, 2002, vol. 4, pp. 4072-4075.
[4] N. Poh, S. Bengio, and J. Korczak, A multi-sample multi-source model for biometric authentication *in Proc. of the IEEE 12th Workshop on Neural Networks for Signal Processing*,, 2002, pp. 375-384.
[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*, 2004, vol. 36(1), pp. 47-62
[6] R. Brunelli, D. Falavigna, Personal Identification using Multiple Cue, *IEEE Trans. on Pattern Analysis and Machine Intelligence*, Vol. 17(10) 1995, 955–966.
[7] J. Suutala, J. Roning, Combining Classifier with Different Footstep Feature Sets and Multiple Samples for Person Identification, *in Proc. of International Conference on Acoustics, Speech and Signal Processing*, 2005, pp. 357-360.
[8] J. Kittler, M. Hatef, R.P.W. Duin, and J. Matas, On combining classifiers, *in Proc. of the IEEE Trans. On Pattern Analysis and Machine Intelligence.*, March 1998, vol. 20, no. 3, pp. 226-239.
[9] M.C. Cheung, M.W. Mak, and S.Y. Kung, Multi-Sample Data-Dependent Fusion of Sorted Score Sequences for Biometric verification, *in Proc. of the IEEE Conference on Acoustics Speech and Signal Processing (ICASSP04)*, Montreal, Canada, 2004, pp. 229-232.
[10] M.C. Cheung, K.K. Yiu, M.W. Mak, and S.Y. Kung, Multi-Sample Fusion with Contrained Feature Transformation for Robust Speaker Verification, *in Proc. of the IEEE Conference on Acoustics Speech and Signal Processing (ICASSP04)*, Montreal, Canada, 2004, pp. 1813-1816.
[11] L.I. Kuncheva, A theoretical Study on Six Classifier Fusion Strategies, *in Proc. of the IEEE Trans. On Pattern Analysis and Machine Intelligence.*, 2001, pp. 348-353.
[12] S.A. Samad, D.A. Ramli, A. Hussain, Person Identification using Lip Motion Sequence, *in Springer-Verlag Berlin Heidelberg*, Part 1, LNAI4692, 2007, pp.839-846.
[13] S.A. Samad, D.A. Ramli, A. Hussain, Lower Face Verification Centered on Lips using Correlation Filters, *in Information Technology Journal*, vol. 6(8), 2007, pp. 1146-1151.
[14] D.A. Ramli, S.A. Samad, A. Hussain, Preprocessing Technique to Voice-Print Analysis for Speaker Recognition, *in Proc. of IEEE Student Conference on Research and Development*, 2007. In press.
[15]http://cslu.cse.ogi.edu/tutordemo/spectrogramReading/spectrogram.html.
[16] R.L. Klevents, R.D. Rodman, *Voice Recognition*, Artech House, INC, 1997.
[17] M. Savvides, K. Venkataramani, and B.V.K. Vijaya Kumar, Incremental Updating of Advanced Correlation Filters for Biometric Authentication System, *in Proc. of ICME*, 2003, vol. 3, pp. 229-232.
[18]B.V.K. Vijaya Kumar, M. Savvides, K. Venkataramani, and C. Xie, Spatial Frequency Domain Image Processing for Biometric Recognition, *in Proc. of International. Conference on Image Processing (ICIP)*, Rochester, NY, 2002, vol. 1, pp. 53-56.
[19] C. Sanderson, and K.K. Paliwal, Noise Compensation in a Multi-Modal Verification System, *in Proc. of International Conference on Acoustics, Speech and Signal Processing*, 2001, pp. 157-160.