Australian Government
Department of Defence
Science and Technology

Face and Voice Fusion for Human Recognition in Non-controlled Environments

Sau Yee Yiu, Dmitri Kamenetsky,
Jason Littlefield and Jonathan Willmore

National Security and ISR Division
Defence Science and Technology Group

DST-Group-TR-3426

ABSTRACT

The individual performance of biometric technologies such as speaker recognition (SR) and face recognition (FR) has enabled their prolific use in applications worldwide (e.g. FR at airports and SR for access to telephone banking and taxation purposes). However, in challenging environments (e.g. CCTV videos), where the data is of low quality, establishing the identity of non-cooperative individuals is still a difficult task.

This paper documents the verification performance gains possible when fusing low quality face and voice samples at the matching score level. Three normalisation and five classifier-based fusion techniques were evaluated on a real life audio-video dataset ('Mobio'). When compared to the performance of the individual biometrics, all fused results showed a notable improvement.

RELEASE LIMITATION
Approved for public release.

Produced by

National Security and ISR Division
Defence Science and Technology Group
PO Box 1500
Edinburgh SA 5111

Telephone: 1300 333 362

November 2017
AR-017-022

Face and Voice Fusion for Human Recognition in Non-controlled Environments

Executive Summary

The individual performance of biometric technologies such as speaker recognition (SR) and face recognition (FR) has enabled their prolific use in applications worldwide (e.g. FR at airports and SR for access to telephone banking and taxation purposes). However, in challenging environments (e.g. surveillance and online videos), where the data is of low quality, establishing the identity of non-cooperative individuals is still a difficult task.

A wide range of approaches for combining biometric samples have been published to overcome some of the practical problems of using only a single biometric trait, demonstrating the benefits of combining different biometric features using fusion algorithms. The Defence Science and Technology (DST) Group has also developed a score-level fusion method (hereinafter referred to as the DST-developed canonical method) [1], which was shown to improve the quality of FR algorithms.

This paper uses a real life non-controlled scenario to examine verification performance gains possible when fusing low quality face and voice samples at the matching score level. Three normalisation methods (*z-score, min-max* and *DST-developed canonical*) and five fusion techniques (*weighed sum*, *Support Vector Machines (SVM) with linear or quadratics kernels* and *Boosting using AdaBoost or RUSBoost algorithms*) were evaluated. The audio-video dataset ('Mobio') used in this evaluation involved both face and voice samples of a person using a mobile phone, which mimics a real life scenario used for authentication.

When compared to the performance of the individual biometrics, the fused results showed a notable improvement. For the Mobio dataset (676 genuine and 48,594 impostor comparisons), at a false match rate (FMR) of 0.1%, the speaker recognition alone achieves a false non-match rate (FNMR) of 35.1%, while FR alone achieves a FNMR of 19.3%. When these two modalities are normalised and fused using the methods listed above, the FNMR is reduced to 7.49-13.9%.

Future research from this evaluation could include:

1. Verifying the significance of current fusion performance using a different database that contains poorer quality face and voice samples than MOBIO.
2. Expanding the current face and voice fusion work to include other emergent biometrics, such as 3D face and body part measurements.
3. Exploring the use of other fusion techniques such as sensor level or feature level fusion.

This page is intentionally blank.

Authors

Sau Yee Yiu
National Security and ISR Division

Sau Yee completed a Bachelor of Applied Science (Mathematical and Computer Modelling) from University of South Australia in 2004. Since then, she has worked as a biometric systems analyst at DST Group, primarily focusing on evaluating the technical performance of biometric systems in various operational environments. Sau Yee also has a high-level focus on the mathematical side that underpins biometrics system performance. Currently her focus is on multi-modal fusion.

Dmitri Kamenetsky
National Security and ISR Division

Dmitri obtained a Bachelor of Science with first class Honours in Computer Science from University of Tasmania in 2005. He then completed a PhD in statistical machine learning at the Australian National University and NICTA in 2009. His PhD thesis investigated methods for inference and parameter estimation in graphical models, in particular the Ising model. In 2010, he joined the NSI Division of the Defence Science and Technology Group. Since then he has worked on object detection, multi-camera pedestrian tracking, atmospheric turbulence mitigation and score-level fusion.

Jason Littlefield
National Security and ISR Division

Jason has been working at DST Group since 2002 where he started as a speech technology developer. Since then his fields of research have included speaker-dependent and speaker-independent automatic speech recognition, text-to-speech synthesis, spoken dialogue systems, and distributed team collaboration platforms. Currently his focus is on speech analytics.

Jonathan Willmore
National Security and ISR Division

Jonathan obtained a Bachelor of Science (Physics and Computer Science) from the University of Adelaide in 1985. Since joining DST Group in 1990, he has worked on several software development projects, automation of processes in military intelligence planning and more recently speech analytics. He has worked in the field of speech processing since 1998 with a special interest in speaker recognition in an operational environment.

Contents

This page is intentionally blank

Glossary

AdaBoost	Adaptive Boosting algorithm
AUC	Area Under Curve
CCTV	Closed-Circuit Television
DET	Detection Error Trade-off
DST Group	Defence Science and Technology Group
FMR	False Match Rate
FNMR	False Non-Match Rate
FR	Face Recognition
FRS	Face Recognition System
ROC	Receiver Operating Characteristic
RUSBoost	Random Under Sampling Boosting algorithm
SR	Speaker Recognition
SVM	Support Vector Machine

This page is intentionally blank

1. Introduction

A biometric system is a pattern recognition system that identifies or verifies a person based on specific physiological or behavioral characteristics that the person possesses [2]. It can establish the identity of an individual based on physical characteristics, as opposed to what they know such as passwords or pins. A biometric identification system compares a biometric sample to all identity profiles in a biometric database, whilst biometric verification compares a biometric sample to a specific single identity profile in a biometric database.

When used for person identity verification (1:1 matching), the output results of a biometric system include a *match score* and a biometric decision of *Genuine* or *Imposter*. If a match score is higher than the threshold, the biometric system provides a *Genuine* decision, otherwise it provides an *Imposter* decision. The performance of a biometric system is typically measured by the accuracy of its detection decisions, in particular its decision making errors.

There are several physiological and behavioral characteristics that can be used for biometric recognition including fingerprints, face, voice, iris, hand and gait [3]. Every biometric has its strengths and weaknesses and its usefulness for an application depends upon aspects such as its uniqueness, collectability, acceptability, and resistance to spoofing. The degree of cooperation from individuals is a significant factor that affects the type and quality of biometric that can be collected. Two examples of high quality biometrics are fingerprints and irises due to their physiological uniqueness amongst individuals, but their collection requires a high degree of cooperation because they are intrusive in terms of time and very close range to the sensor.

In *non-cooperative* or *non-controlled environments* (e.g. CCTV surveillance and online videos), collecting high quality biometrics can be very difficult due to the individual's short time under surveillance and long range from the sensor. In these situations only low quality biometrics from the audio and video are collected, which degrades the performance of biometric systems that use only a single trait of the person. For example, with face recognition (FR), unsatisfactory lighting conditions, occlusions or variations in pose can cause a significant decrease in accuracy [4, 5]. Similarly for speaker recognition (SR), the speech audio signal can be detrimentally affected by channel distortion, environmental background noise or short duration of the speech.

A wide range of approaches for combining biometric samples have been published to overcome some of the practical problems of a single biometric trait, demonstrating the effects of the combination of different biometric features and fusion algorithms [6]. Fusion in biometrics can be performed either at the sensor, feature or score levels, and one strategy designed to improve person identity recognition performance is intra-class multiple sensor fusion [7]. A fusion algorithm developed by the Defence Science and Technology (DST) Group has been shown to improve the performance of a person verification by fusing multiple FR algorithm outputs (hereinafter referred to as the DST-developed canonical method) [1, 8]. Another strategy is inter-class multiple sensor fusion by collecting and fusing multiple low quality biometrics such as face, speech, gait, and body part measurements. Many authors have published research on multimodal biometric systems developed by combining biometric samples from two or more sources [9-11]. Various normalisation techniques needed to transform the raw scores of different biometric modalities into a common domain were also examined in these papers.

In this study, the effectiveness of three normalisation techniques and five score-level fusion approaches in improving the performance of identity verification in non-controlled environments were examined using the face and speech data collected from the inter-class Mobio (audio and video) database [12]. This work focused primarily on score level fusion because this type of fusion only requires the combination of match scores from two or more modalities and no additional information (e.g. feature vectors) is needed from the individual biometric systems. Although score-level fusion discards potentially important information (such as feature sets used by different biometric modalities), the results obtained using this type of fusion have been also shown to be superior [7, 8].

2. Biometrics

2.1 Speaker Recognition

Speaker recognition is a biometric technology that uses features extracted from speech audio to identify or verify a person's identity [13, 14]. An automatic speaker recognition system can either be text-dependent[1] or text-independent[2]. A SR system generally consists of two phases: *enrolment* and *recognition.*

In the *enrolment* phase, a speech sample or "utterance" of a person is recorded to enable the features which characterise the person's voice can be extracted [13-15]. These features are generally represented as a *voice print*, *template*, or *model* which consists of sets of numerical descriptors or feature vectors. Mel-Frequency Cepstral Coefficients (MFCC), Linear Predictive Coding (LPC) and Linear Predictive Cepstral Coefficient (LPCC) are some of the methods that are commonly used to extract the voice features from a speaker's speech sample [13-15].

During the *recognition* phase, the speaker model created during the *enrolment* phase is then used to identify the speaker or verify their identity. Support Vector Machines (SVM), Gaussian Mixture Models (GMMs), Hidden Markov Models (HMMs) and neural networks are the most common modelling techniques that are used for comparing a speech utterance to a previously created speaker model. More recently, an i-vector approach has become widely used in state-of-the-art speaker recognition systems due to its superior discriminating capability [16].

Similar to face recognition (Section 2.2), a match score that represents the maximum likelihood of a match is usually output for each speaker model comparison.

2.2 Face Recognition

Face recognition is one of the primary biometric technologies. It uses the faces present in images and videos to automatically identify a person or verify a person's claimed identity. Traditionally, a five step process is involved in the automated FRS: image acquisition, face detection and alignment, feature extraction, feature matching and declaration of matches. Numerous face recognition algorithms have been proposed which can be categorised as Appearance-based method [13, 14]. Appearance-base methods employ dimensionality reduction technique to represent the whole face of a person in a low-dimensional space for facial comparison.

[1] In a text-dependent SR system, in addition to voice print matching, the text of the speech samples is also compared at the recognition phase (i.e., the speech used for the verification/identification must be the same as what was said at the enrolment).

[2] In a text-independent SR system, the text in the speech sample used for enrolment can be different to the ones used for recognition. This system is often used for identification applications as they require minimal or no cooperation by the speaker.

Acquisition is the first step of the process where a single facial image or a sequence of facial images (i.e., video) of a person are acquired and submitted to the FRS. The image source can either be acquired in real-time or by submitting the person's existing photo to the FRS. After the image is presented to the FRS, the system will first perform *face detection* to localise the face and its facial components from the image. The face is then normalised (*face alignment*) geometrically and photometrically to enable good facial features to be extracted during the *feature extraction* stage. Methods such as Principle Component Analysis (PCA), Gabor wavelets [13, 14] and the latest Convolution Neural Networks (CNN) developments [17] are employed to extract and store (as a template) the unique features of a person's face.

For *feature matching*, the template generated in the *feature extraction* stage is either compared against those templates generated for a database of known faces (*identification*) or to an alternative template of the claimed identity (*verification*). A score which represents the probability of a match is usually output for each comparison. The higher the score the more likely that the comparison is declared a match.

3. Fusion

Fusion is a technique used to merge the results of two or more independent classifiers to obtain a stronger classifier. In this context, a classifier is often referred to as a matcher or an algorithm that classifies a person's biometric traits into one of the two classes: *Genuine* (true match) or *Imposter* (false match) based on a decision boundary. Fusion can either be performed at the sensor or feature level (pre-classification) or at the matching score level (post-classification) [7]. This document focuses primarily on post-classification techniques. A number of algorithms have been proposed for score-level fusion, including Support Vector Machine (SVM) [9, 18] and Boosting [19, 20] based fusion approaches.

3.1 Normalisation

Post-classification fusion approaches often begin with a normalisation step [7, 9], where classifier scores are transformed into a way that assists the classification stage that follows. Sections 3.1.1 to 3.1.4 summarise four commonly used normalisation methods and an in-house method developed by DST Group. It is assumed that the input to each normalisation method is a list of N real-valued scores **S**, where S_k is the k-th score.

3.1.1 Z-Score

Z-score normalisation transforms the data by subtracting its mean μ and dividing by its standard deviation σ:

$$S'_k = \frac{S_k - \mu}{\sigma}.$$

This transformation ensures that the normalised data will have a mean of 0. A positive normalised score indicates a datum above the mean, while a negative normalised score indicates a datum below the mean. If the original data is Gaussian distributed then the normalised data will follow the standard normal distribution with 0 mean and standard deviation of 1. Z-score normalisation does not guarantee a common numerical range for the normalised scores. For example, if the input scores are not Gaussian distributed, z-score normalisation does not retain the input distribution [7].

3.1.2 Min-Max

Min-Max normalisation transforms the data by subtracting its minimum value (min) and dividing by its range:

$$S'_k = \frac{S_k - min}{max - min}.$$

This transformation ensures that the data is scaled to the [0, 1] range.

3.1.3 Rank

To perform the rank-based transformation, first the data is sorted from lowest to highest scores. Let R_k be the 1-based rank (i.e., with ranking number starting from 1) of the k-th score, then the transformation is

$$S'_k = \frac{R_k}{N}.$$

Similar to Min-Max transformation, the rank-based transformation ensures that the data is scaled to the (0,1] range. However, a key difference between the two methods is that the data also becomes uniformly distributed, which has some useful properties for classification. Note that this transformation was not examined in the experiments of this report as this has been implemented as part of the DST-developed canonical method (Section 3.1.4) and it is given here for reference only.

3.1.4 DST-developed Canonical

The DST-developed canonical transformation is an extension of the rank-based method. A brief overview of the method is summarised below. For a detailed theoretical justification of this method please see [1].

The method assumes that one is given scores for both genuine (G) and impostor (I) pairs. First the union of G and I scores is computed and the result is then sorted from smallest to largest, denoted as Q. Let G'_k be the number of elements in G that are equal to Q_k. Formally $G'_k = |\{x \in G: x=Q_k\}|$; similarly for I'_k. Finally, let G* be the cumulative sum of G′, formally:

$$G^*_k = \sum_{n=0}^{k} G'_n.$$

I* is defined in a similar fashion. The canonical score S_k for element k is defined as:

$$S_k = \frac{G_k^*}{2|G|} + \frac{I_k^*}{2|I|} - \frac{G_k'}{4|G|} - \frac{I_k'}{4|I|}.$$

Consider an example with scores common to both G and I, where G = {2, 7, 8, 5} and
I = {3, 1, 2, 4}. The sorted union of these scores is Q = {1, 2, 3, 4, 5, 7, 8}.
Now compute G′ and I′:

$$G' = \{0, 1, 0, 0, 1, 1, 1\},$$
$$I' = \{1, 1, 1, 1, 0, 0, 0\}.$$

Now compute the cumulative sums

$$G^* = \{0, 1, 1, 1, 2, 3, 4\},$$
$$I^* = \{1, 2, 3, 4, 4, 4, 4\}.$$

Finally, the canonical scores for Q become:

$$S = \{0.0625, 0.25, 0.4375, 0.5625, 0.6875, 0.8125, 0.9375\}.$$

These canonical scores are then mapped back to the original raw scores.
Hence G = {2, 7, 8, 5} becomes $G_{canonical}$ = {0.25, 0.8125, 0.9375, 0.6875}, while I = {3, 1, 2, 4} becomes $I_{canonical}$ = {0.4375, 0.0625, 0.25, 0.5625}.

A visual example of the DST-developed canonical transformation is shown below. Figure 1 shows an example scatter plot of randomly generated scores for two modalities (metrics). Blue points are genuine scores, while red points are impostor scores. For ease of interpretation, the genuine scores were forced to be higher than the impostor scores, which is a typical scenario in human verification/identification. Figure 2 shows these scores normalised using the DST-developed canonical transformation. The scores for both metrics are now in the range [0, 1], while their relative position is mostly unchanged.

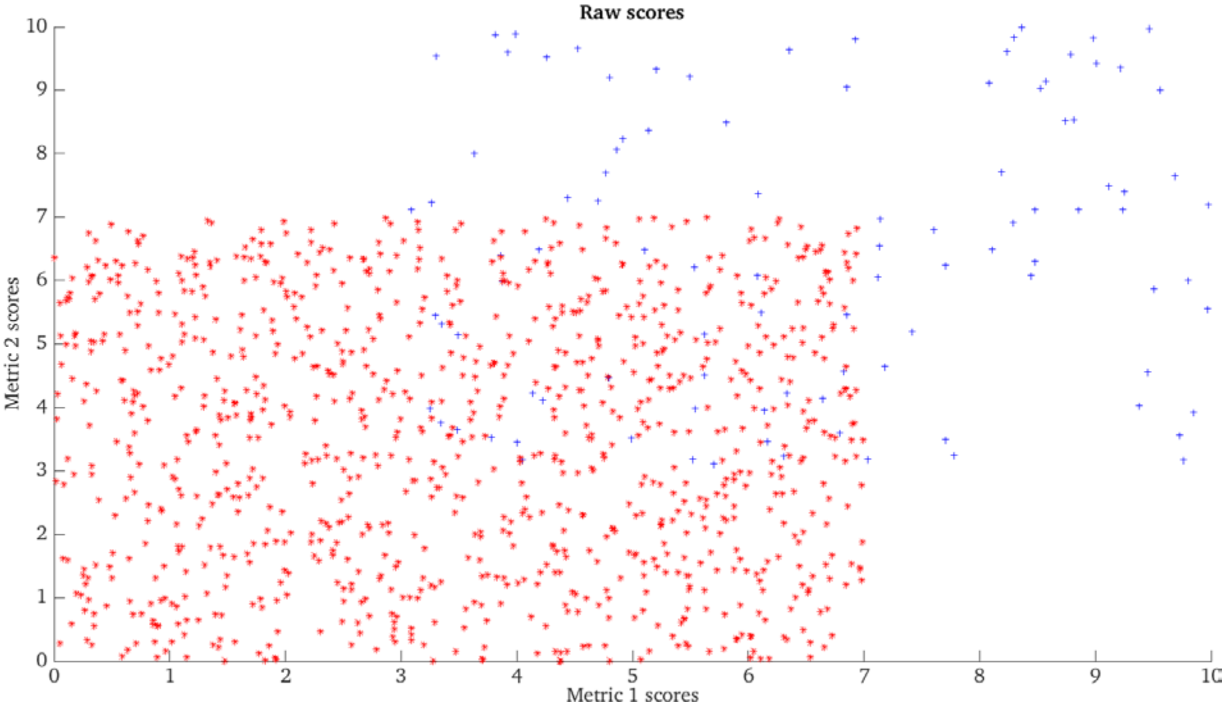

Figure 1: Raw scores for Metric 1 and 2. Blue points are genuine scores, while red points are impostor scores

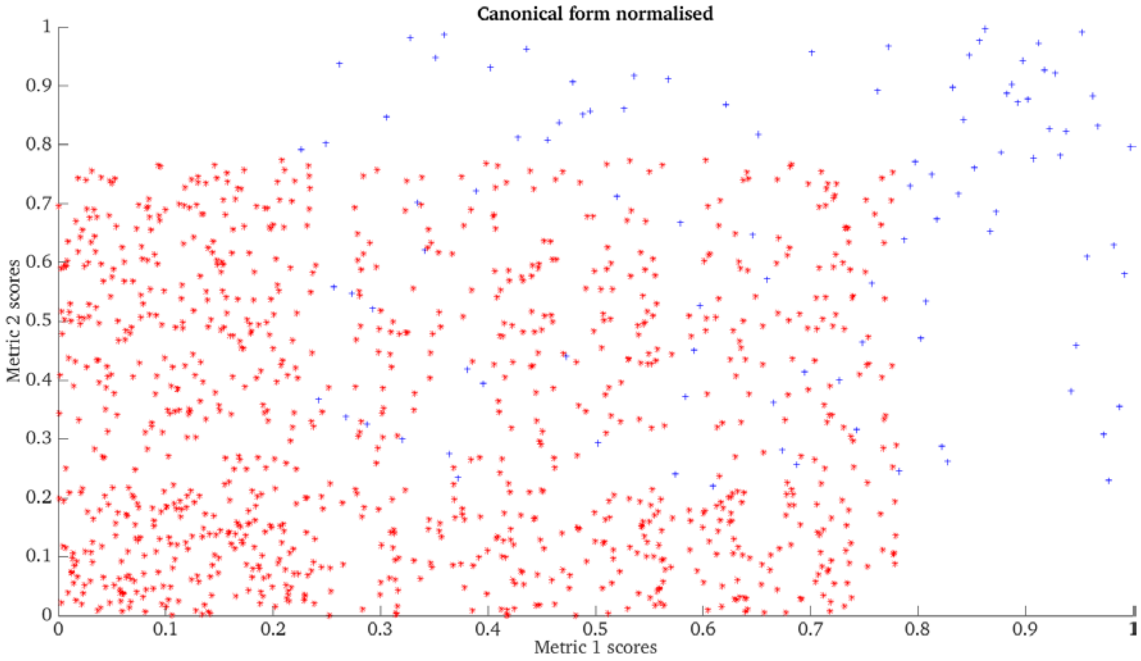

Figure 2: Corresponding scores that have been normalised using the DST-developed canonical transform. Blue points are genuine scores, while red points are impostor scores.

3.2 Weighted Sum Fusion

Weighted sum fusion is the simplest form of score-level fusion. Despite its simplicity, it is efficient and produces reliable results. The method aims to fuse match scores from different modalities by taking their weighted sum.

Although the algorithm works with more than two independent modalities, for simplicity of presentation here only two modalities are used. Let S_1 and S_2 be the score vectors for the two modalities and let w_1 and w_2 be their corresponding weight multipliers. The weighted sum vector W of the two scores is computed as follows:

$$W = w_1 S_1 + w_2 S_2.$$

A Receiver Operating Characteristic (ROC) curve[3] can then be computed by thresholding on W. The performance of the fusion method (for a fixed set of weights w1 and w2) can be measured using the Area Under Curve (AUC) of the ROC [21]. The values of AUC range from 0.5 to 1.0 and a good performing biometric system generally has an AUC value closer to 1.0, whereas, a low AUC indicates poor performance of the system. By iterating over a large set of weights (using brute-force), an optimal weights set that produces the highest AUC can be found.

[3] ROC curves are an accepted method for showing the verification performance of different modalities or algorithms over a range of decision thresholds. The curves plot the false match rate (FMR) on the x-axis against the corresponding true match rate (i.e., 1 – false non-match rate (FNMR)) on the y-axis as a parametric function of the decision threshold.

Consider an example, where there are randomly generated scores for two modalities as shown in Figure 3. Blue points are genuine scores, while red points are impostor scores. For simplicity, fix w_2=1 and vary w_1 only. Figure 4 shows the weighted sum scores when w_1=6.5. Note that this is in fact a one-dimensional figure; the y-coordinates are made different for clarity purposes only. For a given threshold (vertical line at x=44.594) one can compute the false match rate and the true match rate, which contribute to a single point on the ROC curve in Figure 5. The false match rate is the percentage of impostor scores (red) greater than or equal to the threshold, while the true match rate is the percentage of genuine scores (blue) greater than or equal to the threshold. Thus scanning the threshold from left to right in Figure 4 produces the ROC curve in Figure 5. Furthermore each threshold chosen in Figure 4 corresponds to a linear decision boundary in the space of raw scores (Figure 3). In particular, w_1 determines the slope of the decision boundary, while the threshold determines its y-intercept. If w_1 is 0 then the effect of the first modality is "ignored" and so the decision boundary is horizontal. Similarly, if w_2 is 0 (or $w_1 \rightarrow \infty$) then the effect of the second modality is "ignored" and so the decision boundary is vertical. As w_1 increases the slope of the decision boundary decreases. As the threshold increases so does the y-intercept of the decision boundary. Finally, after varying the weight w_1 one arrives at the curve similar to Figure 6. From this figure one can see that the greatest AUC of 0.936 is achieved at w_1=5 and w_2=1. Now these weights can be used during testing of new data.

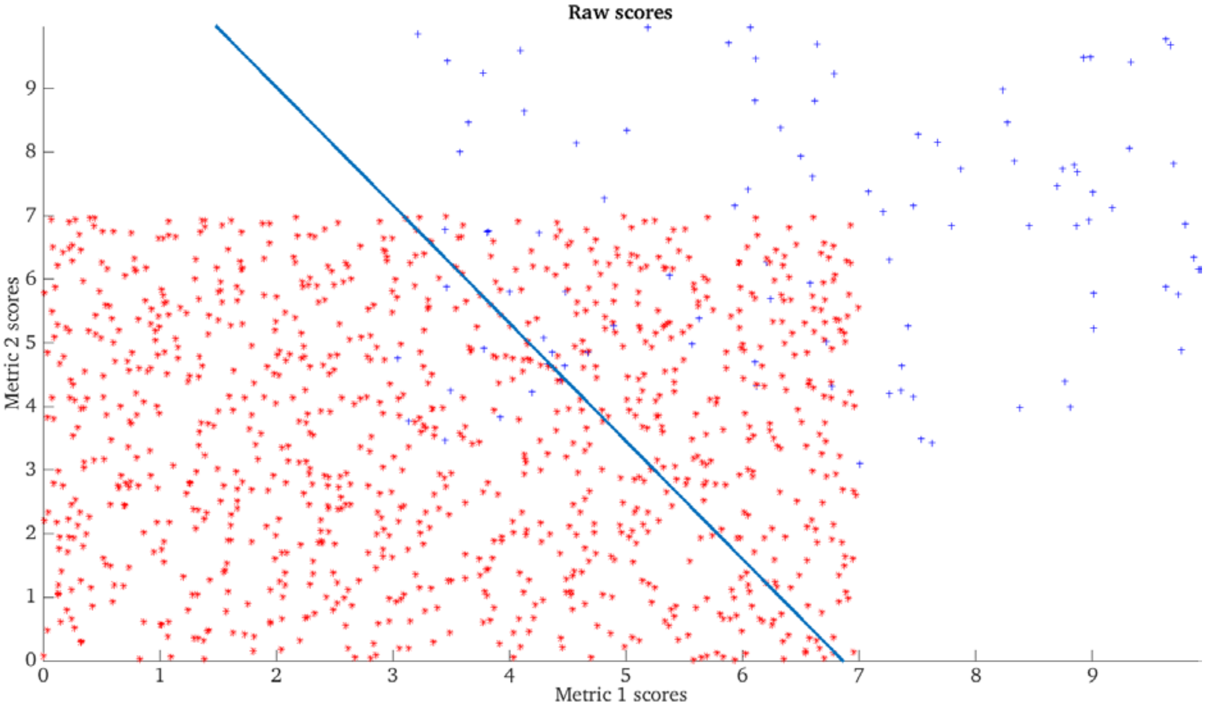

Figure 3: Raw scores with a decision boundary. Blue points are genuine scores, while red points are impostor scores.

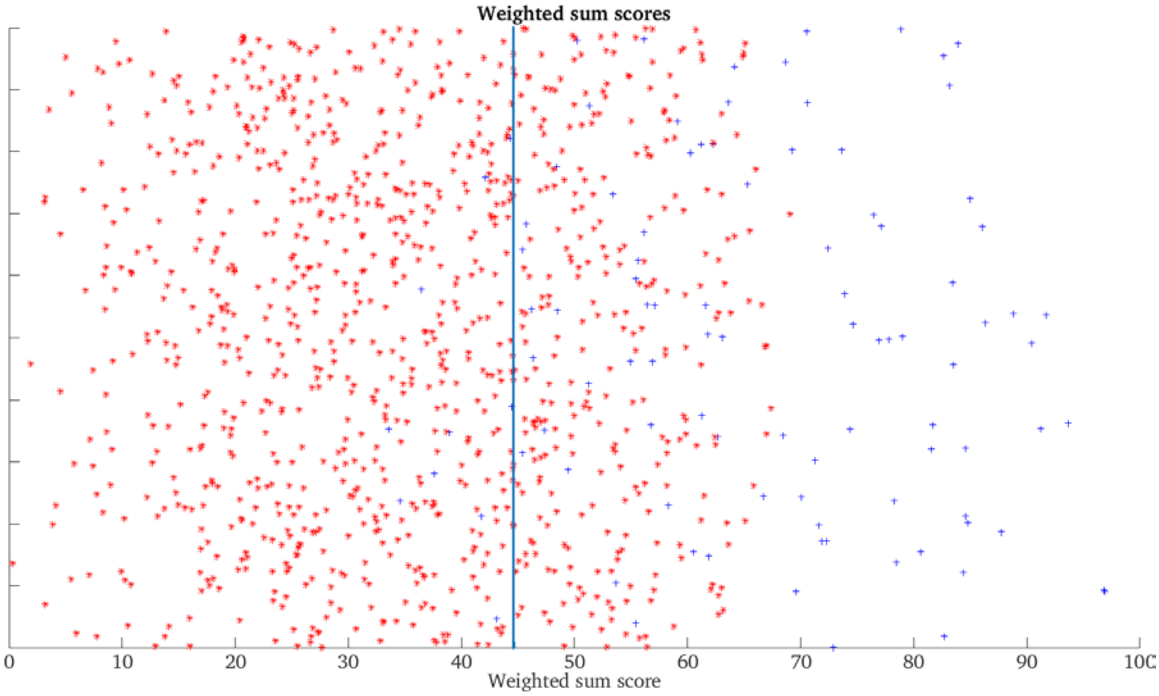

Figure 4: Weighted sum scores with threshold set to 44.594. Blue points are genuine scores, while red points are impostor scores.

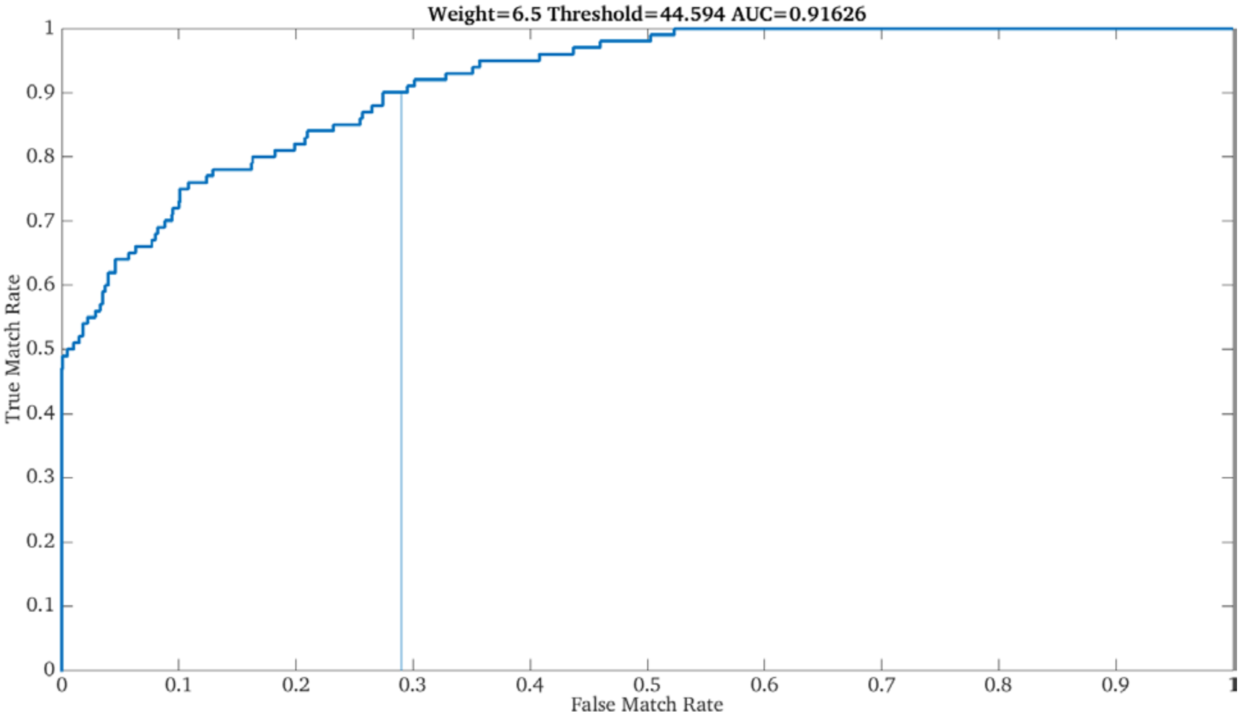

Figure 5: ROC curve resulting from weighted sum fusion. The vertical line indicates the threshold set at 44.594. The area under curve is 0.91626.

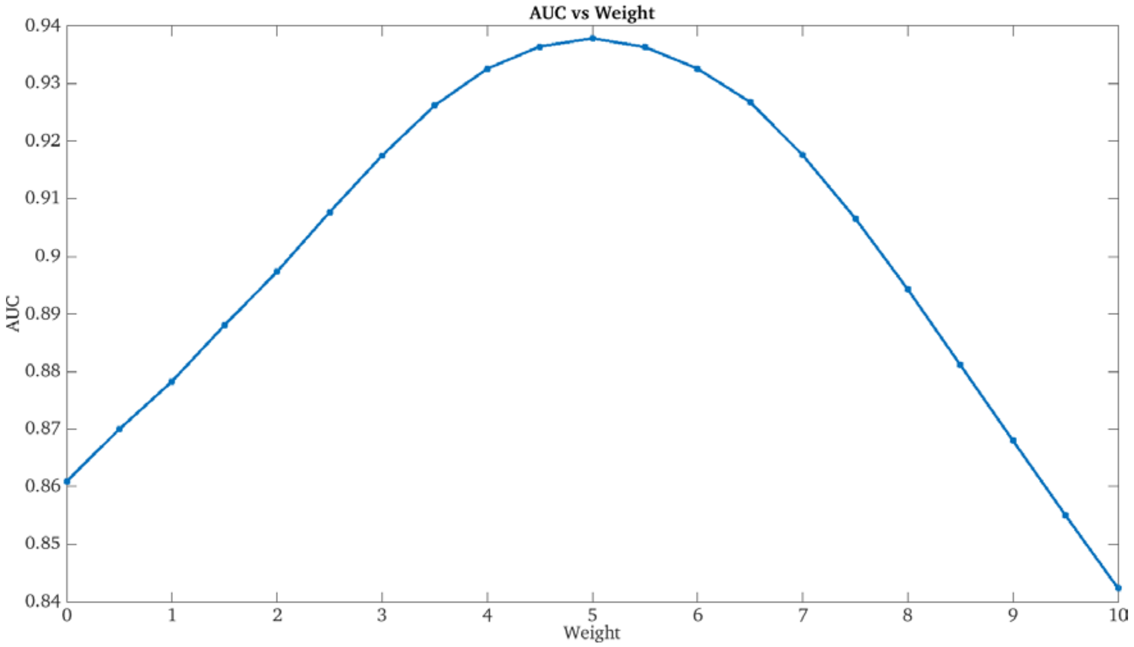

Figure 6: Area under curve as a function of weight w_1. Note that w_2 is fixed at 1.

3.3 Classifier-based Fusion

The weighted sum fusion algorithm described in the previous section can be viewed as a form of a classification with a linear decision boundary. One can also use other popular classifiers for score-level fusion, such as SVMs [22, 23] or Boosting [24, 25].

Boosting algorithms aim to build a strong classifier as a weighted sum of weak learners. For example, the weak learners can be decision tree stumps (one-level decision trees). Such weak learners produce linear separating boundaries (blue line) that are parallel to the axes as seen in Figure 7. In this experiment we used AdaBoost [25] and RUSBoost [26].

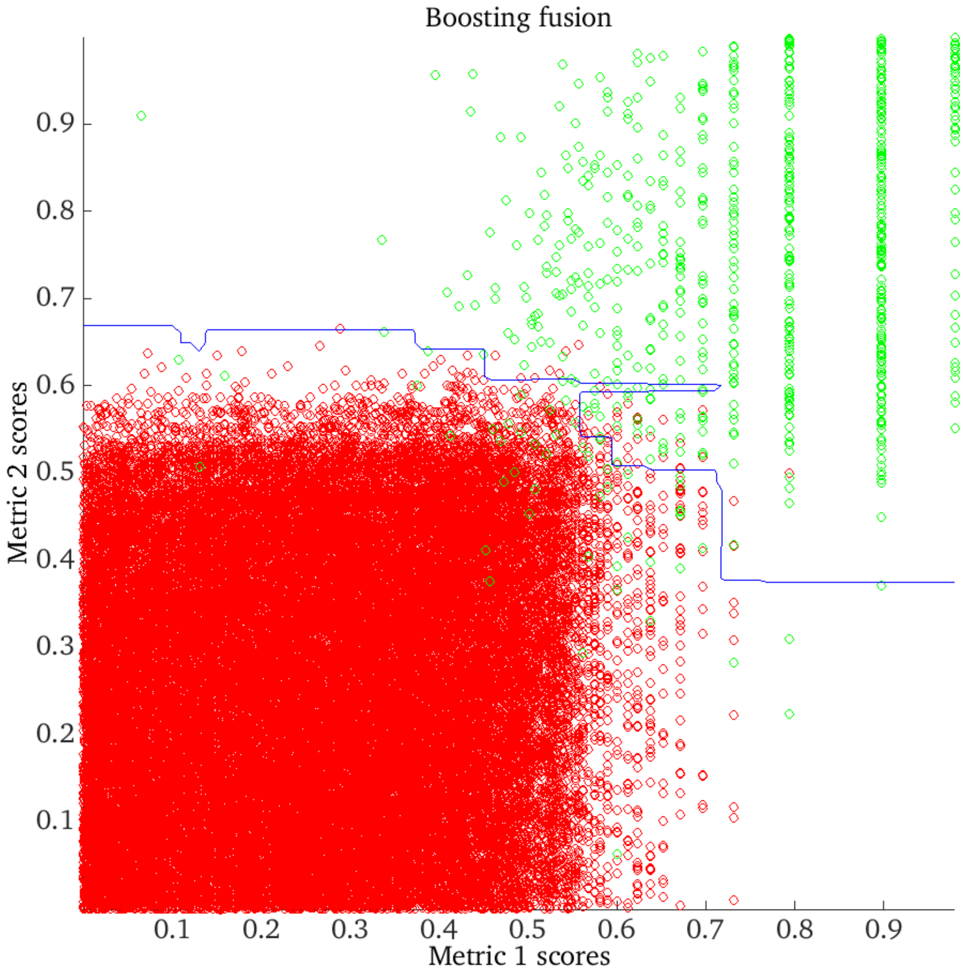

Figure 7: Fusion using Boosting. Red points are imposter scores, while green points are genuine scores. Blue curve indicates the decision boundary.

An SVM is a binary classifier that aims to find the optimal separating boundary between the two classes. For the commonly used max-margin SVMs this is the boundary that produces the largest separation (margin) between the classes. In this report, a linear SVM (Figure 8) and a quadratic SVM (Figure 9) were evaluated, which learn linear and quadratic boundaries of separation, respectively.

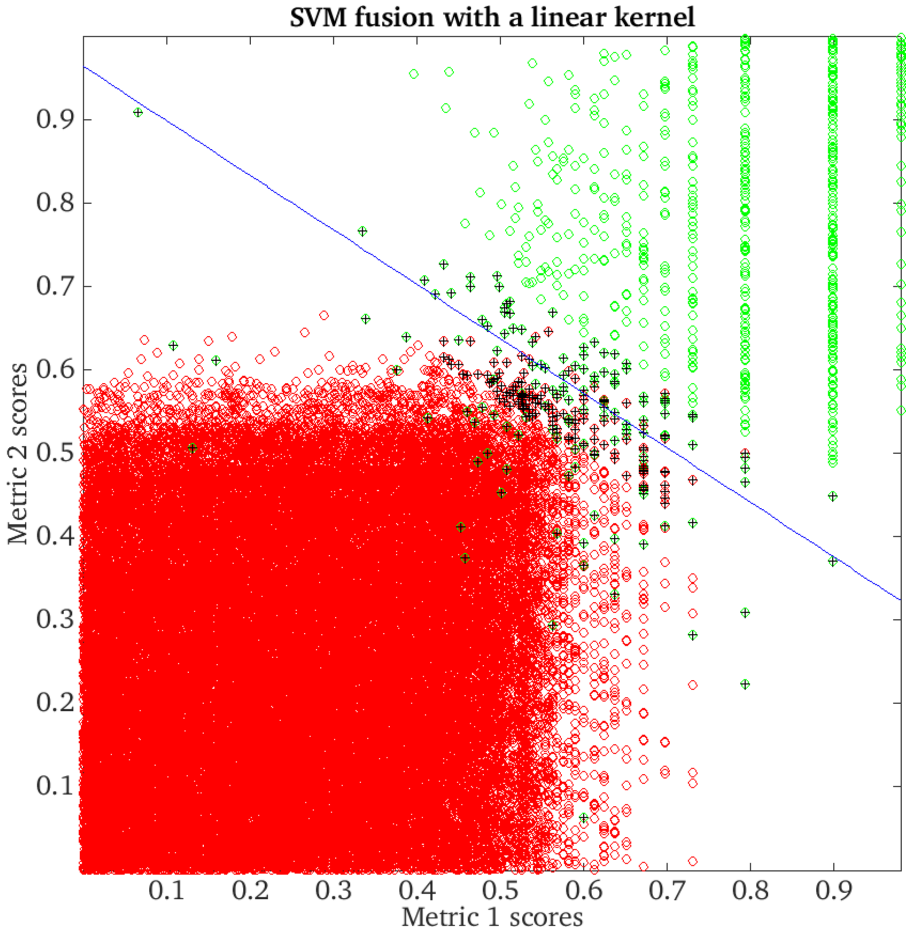

Figure 8: Fusion using a linear SVM. Red points are imposter scores, while green points are genuine scores. Crosses are the support vectors. Blue curve indicates the decision boundary.

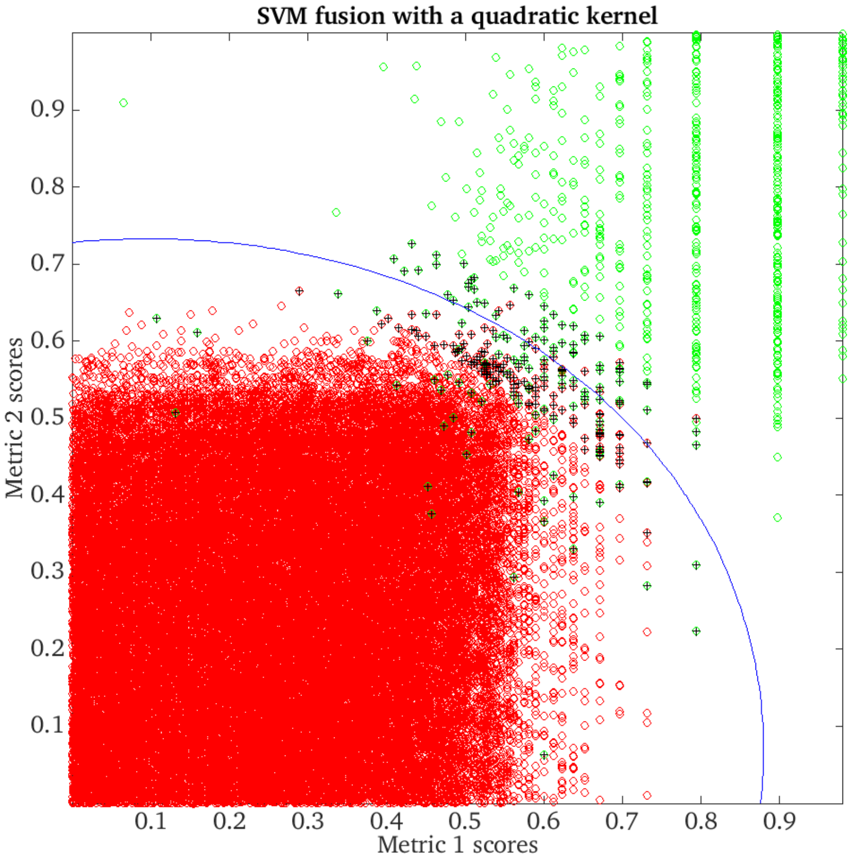

Figure 9: Fusion using a quadratic SVM. Red points are imposter scores, while green points are genuine scores. Crosses are the support vectors. Blue curve indicates the decision boundary.

A learned decision boundary allows one to determine whether a new point should be in one class or another. However, in order to perform fusion one also needs to know the degree of certainty of the above decision. In other words one would like to know the probability P(y=1|x), where x is the input point and y is its label. Unfortunately not all classification algorithms provide such information, but this can be overcome with Platt Scaling [27]. Platt scaling is a method for transforming the outputs of a classification model into a probability distribution over classes. The method works by fitting a logistic regression model to a classifier's scores. It produces probability estimates

$$P(y = 1|x) = \frac{1}{1 + e^{Af(x)+B}},$$

where f(x) are the classifier scores, A and B are two scalar parameters learned by a maximum likelihood method. Now one can use P(y=1|x) to produce ROC curves as was done for W in weighted sum fusion. The advantage of classifier-based fusion over weighted sum fusion is that optimal parameters can be learned, rather than naively found using brute force. On the other hand, weighted sum fusion directly optimises AUC, which is the performance measure one cares about.

4. Experiments

4.1 Algorithms

The face recognition algorithm used for enrolment and verification testing was Cognitec FaceVACS SDK 8.9.5 [28], a popular state-of-art algorithm used in face recognition applications for access control and mobile devices. In terms of speaker recognition, the audio samples used in the experiments were processed through Nuance Identifier ver. 9.4 system [29], a voice biometric system that authenticates users using their own voice.

4.2 Dataset

Phase II of the Idiap Research Institute Mobio dataset [12] was selected for the experiments since it included inter-classbiometric samples of speech and faces captured simultaneously on a mobile phone (Nokia N900). The audio-visual data collected in this dataset approximates the quality of real world data collected in non-controlled environments and included a large enough number of identities for analysing the results with a high degree of confidence.

The Mobio dataset includes MPEG-4 audio video recordings from 150 people (51 females and 99 males) with 6 sessions per person and 11 recordings per session. The recordings captured the participants answering short response questions, pre-defined text read out loud and about 10 seconds of free speech. As highlighted by [12], due to the recordings being captured using a hand held device on different days and at multiple locations there is significant variability in lighting, camera angle and background of images as well as the quality of speech audio. Additionally a large variation in facial expression, hair style, clothing, image sharpness and occlusion were found in these images.

For the experiments, the Mobio dataset was divided up into enrolment and verification testing sets. For each of the 150 people, the longest recording from the first 5 sessions was selected for testing, and the longest recording from the remaining session was selected for enrolment. The frame at the 3 second mark was extracted to create the enrolment and verification testing sets for face recognition. This resulted in an enrolment database of 150 people (1 face image and 1 audio sample per person) and a verification testing database of 150 people x 5 sessions (5 face images and 5 audio samples per person). Thus, 750 genuine and 55 875 impostor comparisons can be obtained from this database[4]. However, due to the quality of the face imagery, some of the people's images failed the enrolment process by the Cognitec FR algorithm and as a result of that images and audio files of these people were excluded from the experiment. So in total, 676 genuine and 48 594 impostor comparisons were used to analyse the system's performance. *Table 1* below provides a breakdown of the comparisons.

[4] An impostor comparison occurs when the person to be verified claims someone else's identity. In this assessment, each of the 5 face images and 5 audio samples of the 150 people are compared to another person's face images and audio samples in the database. This resulted in a total of 55, 875 impostor comparisons (i.e., (150 people x 5 x 149 impostors)/2) per modality.

Table 1 Breakdown of the comparisons for Genuines, Impostors and All Authenticities versus Female to Female, Male to Male, Male to Female and All Genders.

Authenticity	Female to Female	Male to Male	Male to Female	All
Genuines	229(3.9%)	447(2.0%)	0(0.0%)	676(1.4%)
Impostors	5634(96.1%)	21477(98.0%)	21483(100.0%)	48594(98.6%)
All	5863(11.9%)	21924(44.5%)	21483(43.6%)	49270(100.0%)

4.3 Results

In this section, all matching performances were assessed against the standard metrics of False Match Rate (FMR) and False Non-Match Rate (FNMR) [14]. FMR is calculated as the proportion of matches with wrongful claims of identity that are incorrectly accepted by the biometric system, and FNMR represents the proportion of matches with truthful claims of identity that are incorrectly rejected by the system. These two metrics were calculated over a range of thresholds (or match scores) and reported using a cumulative probability plot and/or a detection error trade-off (DET) plot.

4.3.1 Performance of Individual Modalities

The cumulative probability plots showing the FMR and FNMR performance for the face and speech samples of the Mobio dataset are shown in Figure 10 and Figure 11, respectively. As part of the individual modalities performance assessment, the dataset was also divided based on gender to establish any differences in performance between the two gender groups.

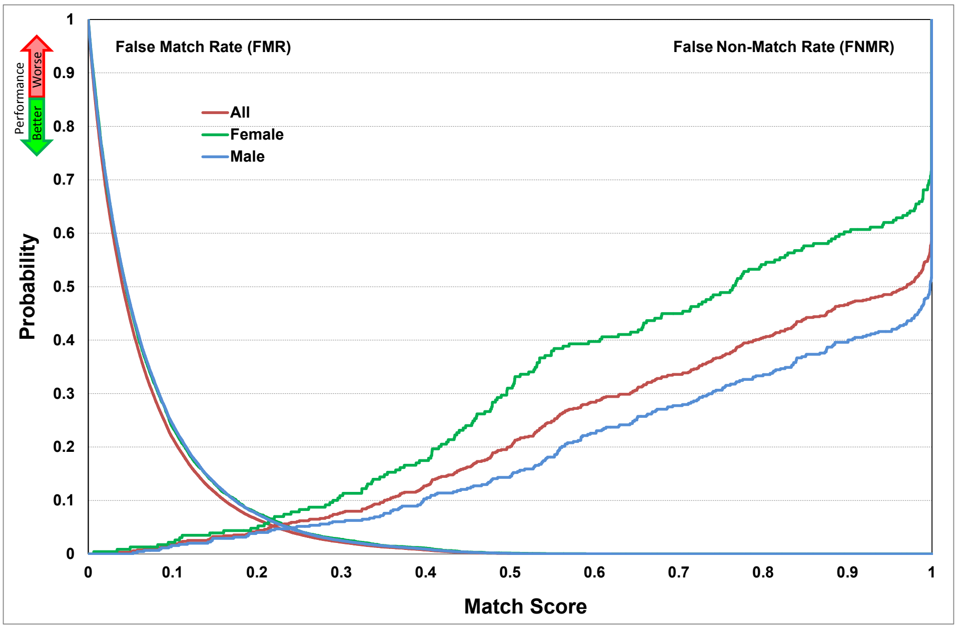

Figure 10: Cumulative Probability Plot – Face Recognition

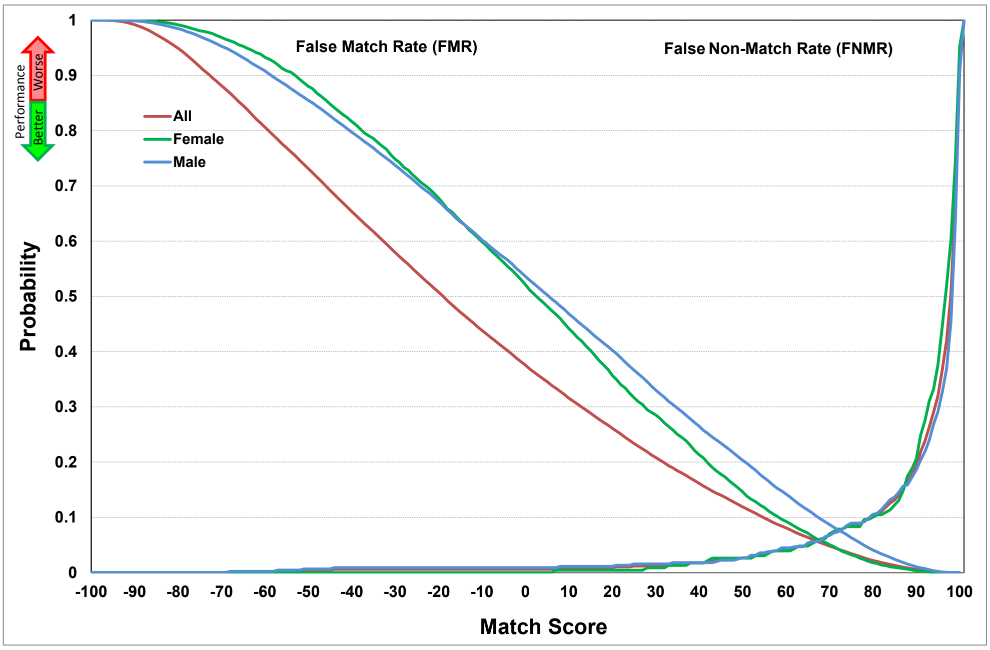

Figure 11: Cumulative Probability Plot – Speaker Recognition

For the correct interpretation of the cumulative probability plot, the better FMR or FNMR performance is at the bottom of the plot and a minimal overlap between FMR and FNMR curves generally indicates a better performing modality or dataset. Using these plots, a threshold could also be set for each modality (and its gender subgroup) such that both false match and false non-match errors can be minimised.

In terms of face recognition performance, Figure 10 shows that there was a large variation between the gender subgroups in terms of FNMR with males having lower FNMR, and, a very minimal variation in terms of FMR. This implies that the FR algorithm recognises males more easily than females and this is consistent with the results reported in previous studies [30].

For speaker recognition, a small variation in FMR was found between the gender subgroups (as shown in Figure 11), and a minimal variation in terms of FNMR. The FMR results indicate that males with a wrongful claim of identity may have a higher chance of deceiving the SR system if the system is set a threshold ≥ 0. A better FMR performance was obtained for the males and females combined (“All”) group because this mimics the zero effort impostor scenario where an individual makes no attempt to increase his/her chance of success to deceive the SR system. In this experiment, an impostor could claim any identity, not just those having the same gender. Since the chance of returning a high match score when comparing two voice samples of different gender is low, a better FMR performance could be achieved for the “All” group (as indicated in Figure 11).

The overall performance of the individual modalities and the effect of gender subgroups is also summarised in a DET plot (Figure 12), where the better performing modality (or gender subgroup) is at the bottom left corner of the plot.

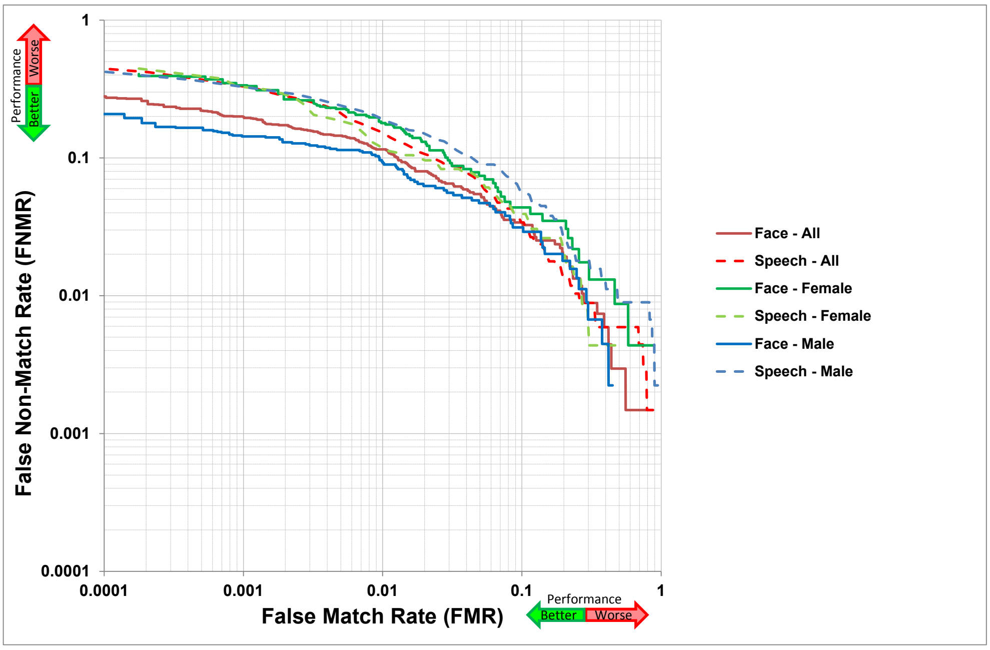

Figure 12: DET Plot – Face Recognition Vs Speaker Recognition

As shown in Figure 12, the FR algorithm appears to offer a better overall matching performance than the speaker recognition system at low FMRs. For the gender subgroups, the performance appears to be better when the face imagery is used as the biometric samples for the males. For the female group, the performance appears to be better when their speech samples are used instead.

4.3.2 Multimodal Fusion Performance

In this section, the match scores from the face and speaker recognition systems were fused directly using the weighted sum method (as described in section 3.2) to demonstrate how the performance of unimodal biometric systems can be improved. The fusion performance in this and subsequent analyses were evaluated by using the area under the Receiver Operating Characteristic (ROC) curve (denoted as AUC).

For the weighted sum method, a search is required to find the optimum weights, that is, the weights that give the maximum AUC. Let **w** be the weight of the face scores, then the weight of the speaker scores is (10-**w)**. In this study, we tested all **w** from 0 to 10 in increments of 0.1. For example, when fusing the face and speaker match scores the weights considered were (10, 0), (9.9, 0.1), (9.8, 0.2), …, (0.1, 9.9), (0, 10) respectively.

Figure 13 shows the weighted sum fusion of face and speaker identities on the whole dataset, as well as, for each gender subgroup. The horizontal axis shows the weight **w** of the speaker recognition system. When **w**=0 the AUC is identical to the FR algorithm alone, and is labelled on the left-hand vertical axis. When **w**=10 the AUC is identical to the speaker recognition system alone, and this is labelled on the right-hand vertical axis.

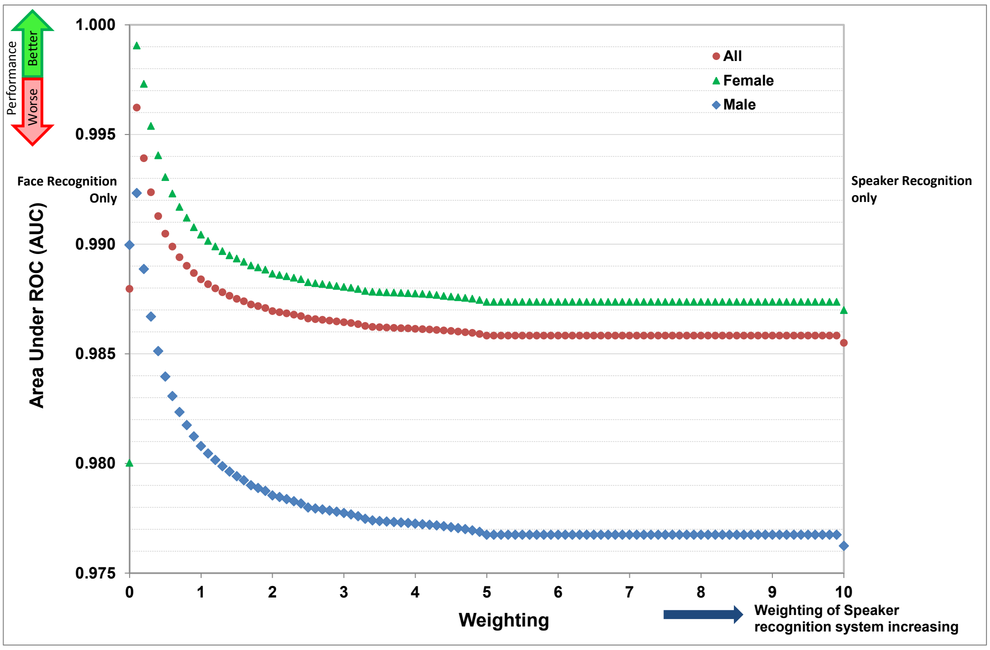

Figure 13: Weighted sum fusion of Face and Speaker Match Scores

It can be seen from Figure 13 that for all groups, the overall performance of the speaker recognition system is improved if it is fused with the FR algorithm, for all weightings examined. The optimum weights that give the maximum AUC (i.e., the best fusion performance obtained) seem to be the same between the whole database and gender subgroups. It should be noted that when this dataset is examined as a combined gender group, the fusion performance appears to be poorer than using the FR algorithm alone if the weighting towards the speech match score is higher than 1.0. However, the fusion performance depends heavily on the choice of the normalisation method and the classification algorithm used. These choices are examined in the next section.

Figure 14 shows the DET plots for the face and speaker recognition algorithms alone (all and gender subgroups), and the DET plots resulting from the best weightings for fusion of the two modalities on the whole dataset and by its gender subgroups. Figure 15 compares the performance of individual modalities with the best fused results at FMR=0.01.

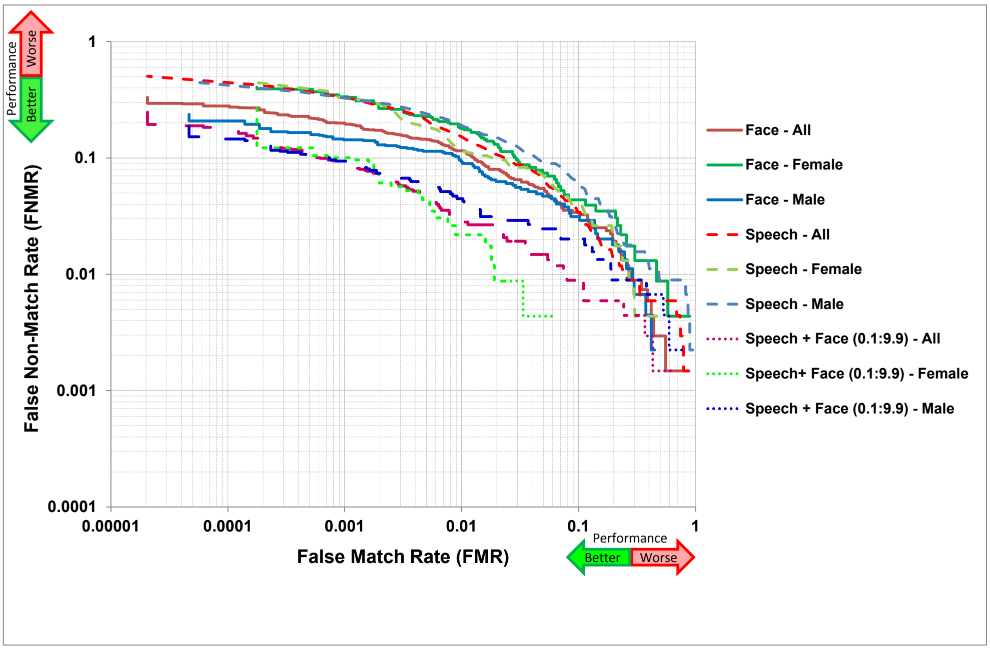

Figure 14: DET plots of individual algorithms and fused algorithms using best weights (all and by gender)

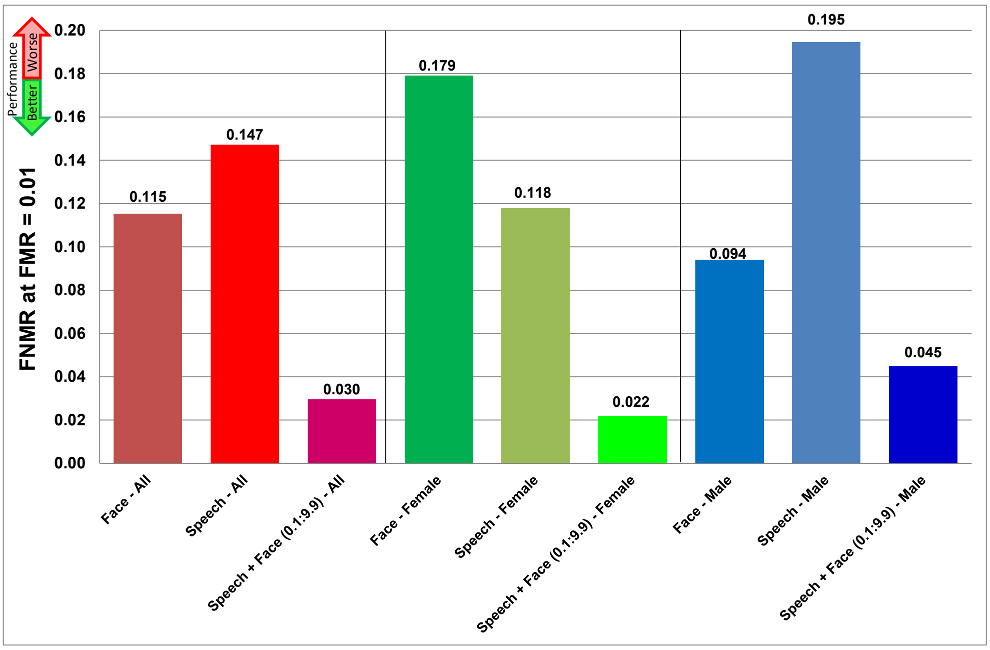

Figure 15: FNMR at FMR = 0.01 (all and by gender)

In this figure, the shorter columns represent better performance. Therefore, at FMR = 0.01, there is a significant improvement in FNMR performance when the two modalities are fused at the score level. This means that the percentage of truthful claims of identity matches that are incorrectly rejected is on average reduced to 2.2 to 4.5% when the two modalities are fused.

4.3.3 Evaluation of Score Normalisation and Classification Techniques

In the previous section, fusion using a simple weighted sum method was shown to provide a drastic improvement in performance over the individual face and speaker recognition systems. To further study the impact of fusion on the individual modality performances, four different normalisations and five different score-level fusion techniques (using classification approaches) were applied to the Mobio dataset. These normalisation and classification-based fusion methods were implemented in Matlab and are described in Sections 3.1 and 3.3, respectively.

A total of 20 different normalisation and classification-based fusion methods were conducted in this part of the evaluation. For each test, a training set was used to train each classification-based fusion method and the resulting trained model (i.e., a model with the optimised parameters) is then used to classify a testing set.[5] [6] It should be noted that in order to obtain the best kernel parameters for the SVM models, a five-fold cross-validation procedure [31] was applied to each SVM classification algorithm.

Table 2 shows the AUC performance of the testing set when the weighted sum (using the best weightings), SVM with linear kernel, SVM with quadratic kernel, RUSBoost [26] and AdaBoost [25] were applied to the raw scores and three types of normalised scores as labelled in Table 2.

[5] Since the combined effect of different normalisation and fusion methods is the major focus of this section, the performances between the gender subgroups will not be examined here.

[6] 70% and 30% of the Mobio dataset were randomly partition into a training set and testing set using stratification such that both sets have approximately the same genuine and impostor proportions as in the original data set.

Table 2 Performance of Test 1 to 20

Test	Normalisation Techniques	Classification-Based Fusion Techniques	AUC
1	Raw Scores	Weighted Sum	0.9989
2	Raw Scores	SVM (linear kernel)	0.9991
3	Raw Scores	SVM (quadratic kernel)	0.9943
4	Raw Score	Boosting (RUSBoost)	0.9983
5	Raw Score	Boosting (AdaBoost)	0.9988
6	Z-score	Weighted Sum	0.9987
7	Z-score	SVM (linear kernel)	0.9991
8	Z-score	SVM (quadratic kernel)	0.9964
9	Z-score	Boosting (RUSBoost)	0.9982
10	Z-score	Boosting (AdaBoost)	0.9989
11	Min-Max	Weighted Sum	0.9987
12	Min-Max	SVM (linear kernel)	0.9991
13	Min-Max	SVM (quadratic kernel)	0.9946
14	Min-Max	Boosting (RUSBoost)	0.9984
15	Min-Max	Boosting (AdaBoost)	0.9988
16	DST- developed Canonical	Weighted Sum	0.9989
17	DST- developed Canonical	SVM (linear kernel)	0.9986
18	DST- developed Canonical	SVM (quadratic kernel)	0.9966
19	DST- developed Canonical	Boosting (RUSBoost)	0.9982
20	DST- developed Canonical	Boosting (AdaBoost)	0.9987

**AUCs for face and speaker recognition using raw match scores are 0.9880 and 0.9855 respectively.*

As shown in *Table 2*, all fusion methods achieved a relatively high and similar performance on the testing set, with AUCs ranging from 0.9943 to 0.9991. For the raw scores and normalised scores using z-score and min-max normalisations, SVM with linear kernel was the best performing classifier (AUC = 0.9991) on the testing set. However, the differences in these results are considered not significant as all classification models achieved a near-optimal AUC.

The DET plots of the individual modalities and the DET plots for the weighted sum (using the best weightings), SVM with linear kernel, SVM with quadratic kernel, RUSBoost and AdaBoost over the raw and normalised testing set are shown in Figure 16 to Figure 20.

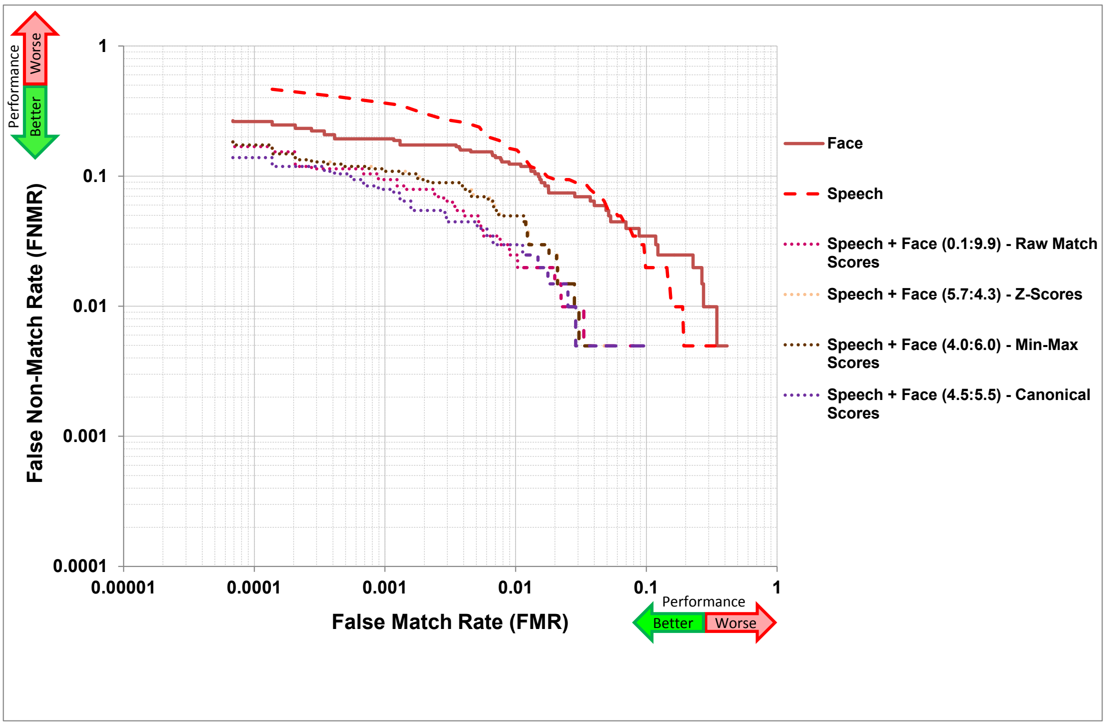

Figure 16: DET plots for Weighted Sum Fusion Method

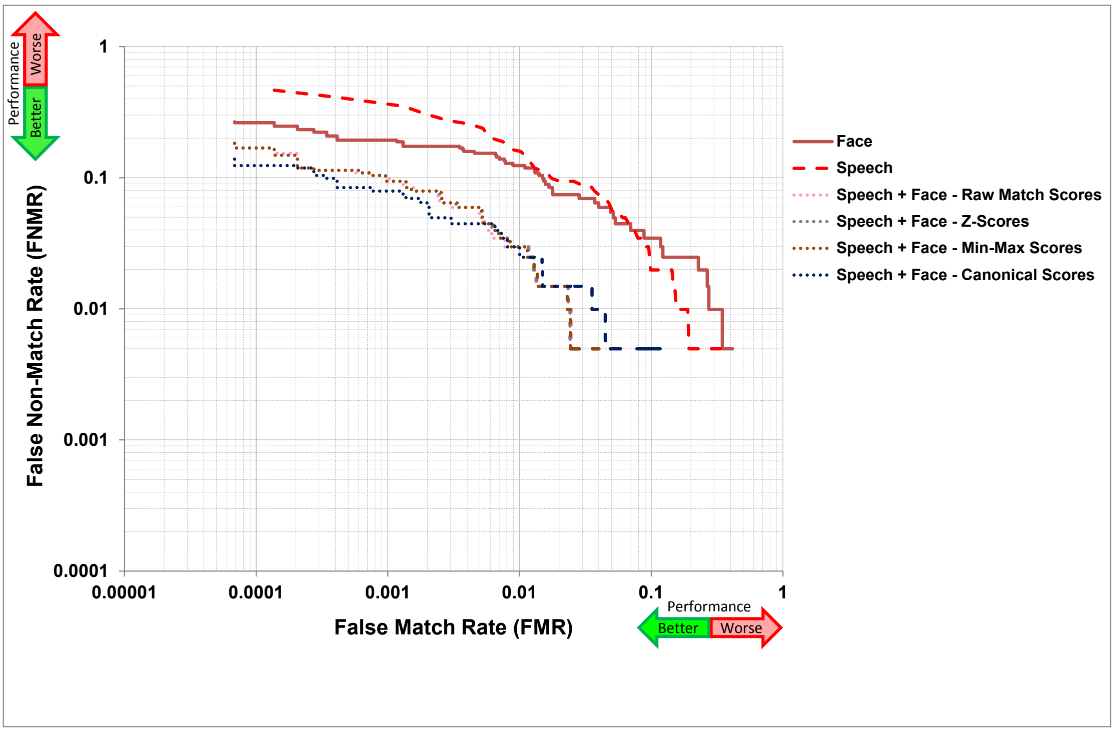

Figure 17: DET plots for SVM-based Fusion Method (with Linear Kernel)

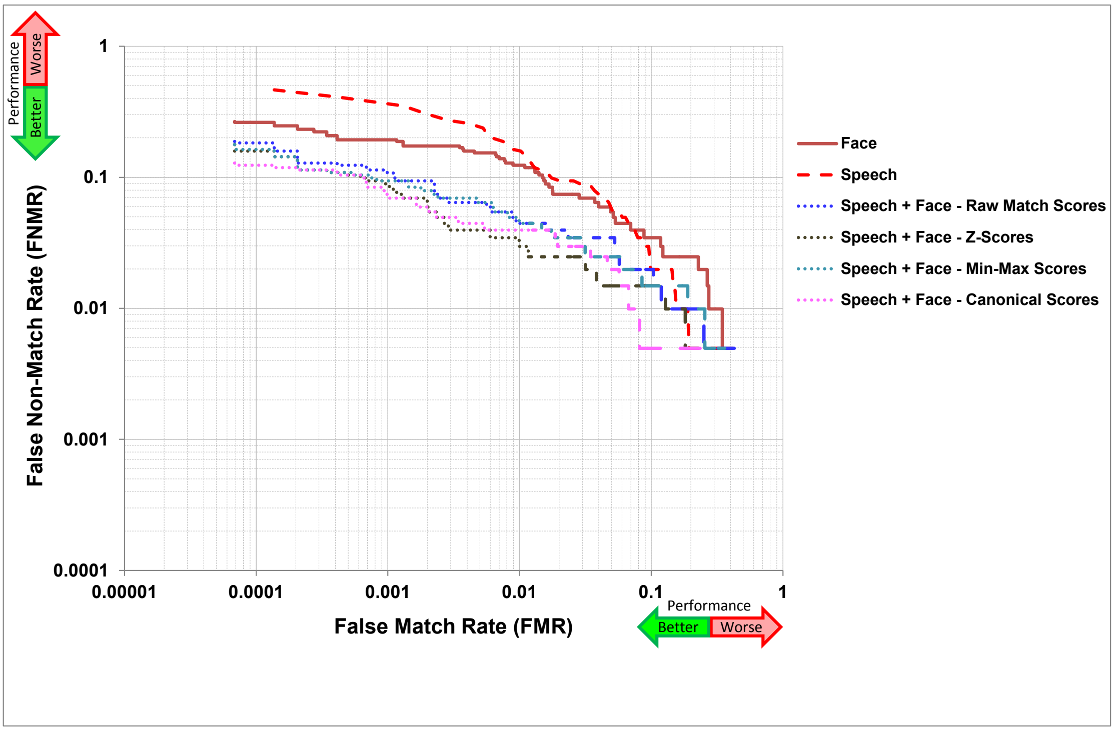

Figure 18: DET plots for SVM-based Fusion Method (with Quadratic Kernel)

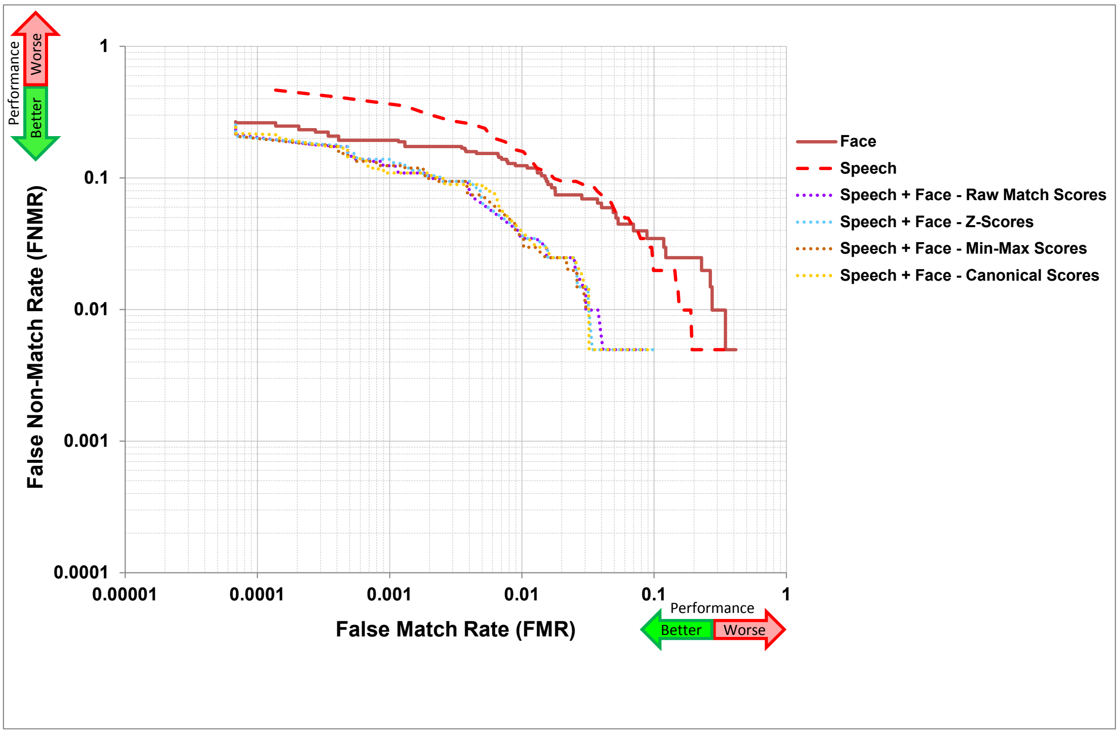

Figure 19: DET plots for Fusion using RUSBoost Algorithm

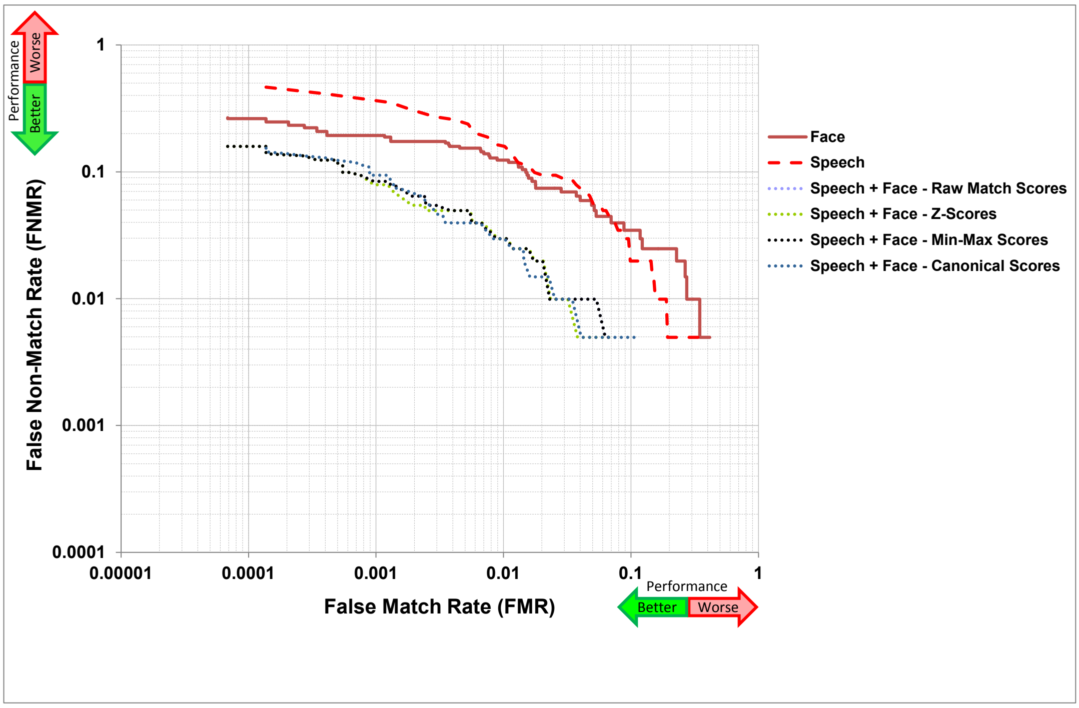

Figure 20: DET plots for Fusion using AdaBoost Algorithm

As can be seen from the DET plots of Figure 16 to Figure 20, fusion using all five classification methods provides better performance than the corresponding unimodal biometric system for both the raw scores and three proposed normalisation methods. For the weighted sum and SVM with linear kernel, normalisation using the DST-developed canonical transformation seems to perform better than Min-Max and Z-score normalisations at low FMRs. However, the performances of the two boosting approaches do not seem to be affected significantly by score normalisations (as indicated in Figure 19 and Figure 20).

For each normalisation and classification technique, the FNMR metric was compared at a FMR of 0.001 and 0.01, which are the typical operating thresholds of a biometric system. These comparisons are shown in Figure 21 and Figure 22 respectively.

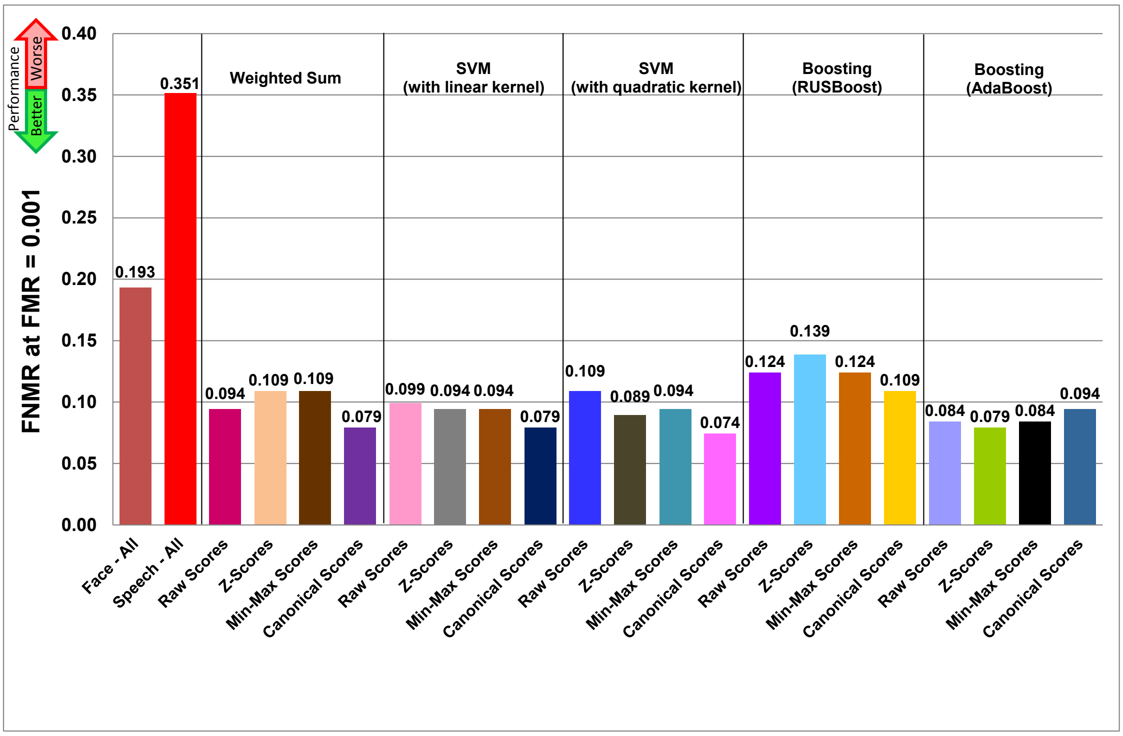

Figure 21: FNMR at FMR = 0.001

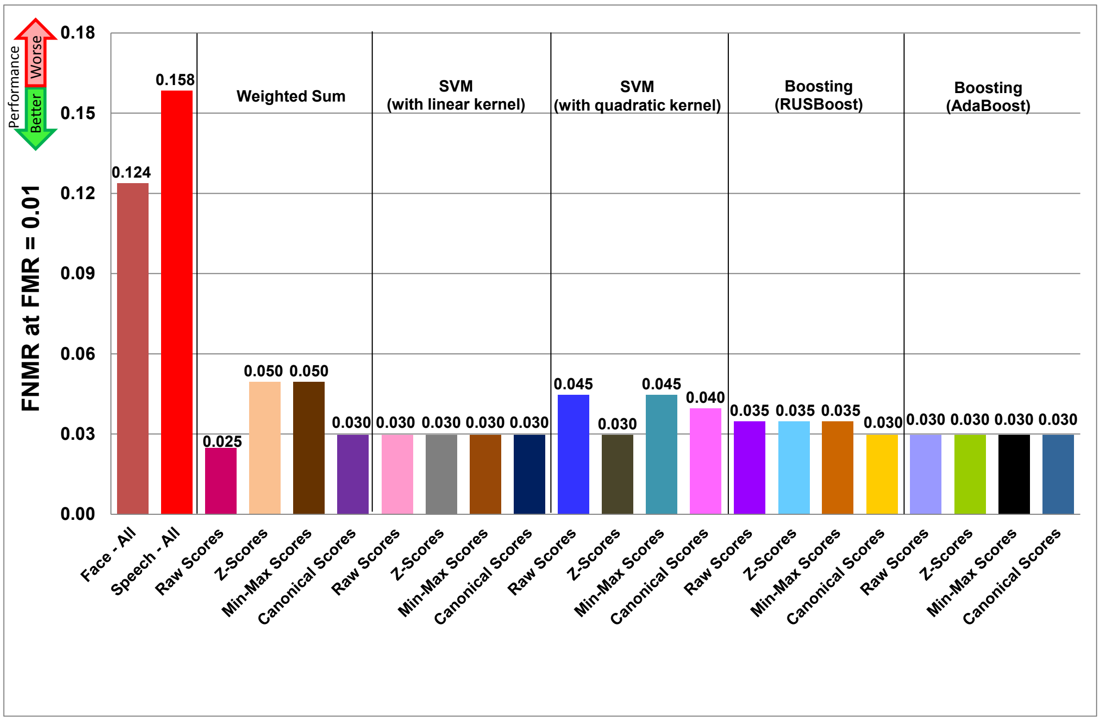

Figure 22: FNMR at FMR = 0.01

At FMR = 0.001, each fusion method using raw scores and any of the three normalised score types outperforms the best single modality performance (i.e., face). The highest performance gain for the canonical transformation was obtained by SVM with quadratic kernel, which led to approximately 11.9% reduction of FNMR at a FMR of 0.1% when compared to the best single modality performance (i.e., face). This result is considered not significant though as all of the classifiers achieved a near-optimal AUC.

For all of the scores normalisation methods, the performance of the SVM with linear kernel is comparable with the performance of the AdaBoost algorithm at a FMR of 1.0%. From Figure 22, it can be seen that sum rule-based fusion just on the raw scores achieved the best performance. The underperformances of sum rule preceded by score normalisations are likely due to different parameters being used to normalise the training and testing datasets. For example, a slightly different mean μ and standard deviation σ were applied to the training and testing sets in z-score normalisation. So when evaluating the sum rule-based fusion performance on the testing set, the best weightings obtained from the training dataset may not be optimal for the testing data and therefore produce performance that is worse than the raw scores.

5. Conclusion and Future Work

This study examines the effect of different normalisation and score level fusion methods on the performance of establishing the identity of individuals in non-controlled environments. The audio-video dataset ('Mobio') used in this study collected both face and voice samples of a person using a mobile phone, which mimics a real life scenario used for authentication.

Results show that improvements in performance were outstanding when either the raw scores or normalised scores of the two modalities were combined using various classifier-based fusion methods. *z-score* and *min-max* normalisation methods followed by a linear SVM based fusion method offer the best fusion performance (AUC = 0.9991). Similar performances were obtained using the *canonical score* (a normalisation method developed by the DST Group [1]). The differences in these results were relatively small but are not significant because all of the classification models achieved a near-optimal AUC.

For the Mobio dataset, at a FMR of 0.01 (1.0%), the speaker recognition alone achieves a FNMR of 0.351 (35.1%), while face recognition alone achieves a FNMR of 0.193 (19.3%). When the scores of the two modalities are normalised and fused using various methods, the FNMR drops down to 7.49-13.9%. Therefore, these results demonstrate that a significant improvement in recognition performance can be achieved using an inter-class biometric system that uses poor quality face and voice samples for user authentication.

Based on the evaluations on the Mobio dataset, a number of future research have been identified for fusing multiple low quality biometric samples to improve establishing the identity of individuals in non-cooperative or non-controlled environments. Some of the potential future research areas are discussed in brief below:

1. The repeatability of current fusion performance would need to be verified using a more challenging dataset that has a higher number of subjects and poorer quality of voice or face samples than MOBIO.
2. Fusion of other emergent biometric modalities such as 3D face, body part measurements and gait.
3. The fusion performance of the weighted sum method could be improved by refining the granularity of the weightings, particularly as the optimum is approached.
4. The current fusion performance evaluation could be expanded to include other pre-classification fusion techniques such as sensor level or feature level fusion.

6. References

1. Bourn, S., Kamenetsky, D., and Yiu, S.Y. (2016) *A formal approach to practical binary classification using a continuous predictor*. DST-Group-TR-XXXX (under review), Edinburgh, Defence Science and Technology Group
2. Fouda, Y. M. (2012) Fusion of Face and Voice: An Improvement. *International Journal of Computer Science and Network Security* **12** (4) 37 - 43
3. Heyer, R. (2008) *Biometrics Technology Review*. DSTO-GD-0538, Edinburgh, Defence Science and Technology Group
4. Hanton, K. (2005) *The Effects of Variable Lighting Conditions on Facial Recognition Scores*. [Honours], Royal Melbourne Institute of Technology, Unpublished Honours Thesis
5. McLindin, B., Bastian, V., Fletcher, K., Johnson, R., Yiu, S. Y., Calic, D. and Hanton, K. (2005) *An Investigation into the Suitability of a Provisional Image Quality Standard for Facial Recognition Implementation*. DSTO-TR-1717, Defence Science and Technology Group
6. Kuncheva, L. I. (2004) *Combining Pattern Classifiers – Methods and Algorithms*, John Wiley & Sons
7. Jain, A. N., K. and Ross, A. (2005) Score Normalisation in Multimodal Biometric Systems. *Journal of Pattern Recognition* **38** 2270 - 2285
8. Yiu, S. Y., Bourn, S., McLindin, B., Malec, C. and Hanton, K. (2015) *Score Level Fusion for 1:1 Facial Algorithms*. Defence Science and Technology Group
9. Horng, S. J., Chen, Y.H., Run, R.S., Chen, R.J., Lai, J.L. and Sentosal, K.O. (2009) An Improved Score Level Fusion in Multimodal Biometric Systems. In: *International Conference on Parallel and Distributed Computing, Applications and Technologies,*
10. Sanderson, C. (2003) *Automatic Person Verification Using Speech and Face Information*. Griffith University
11. Sanderson, C. and Paliwal, K. K. (2004) Identity Verification Using Speech and Face Information. *Digital Signal Processing* **14** (5) 449 - 480
12. McCool, C., Marcel, S., Hadid, A., Pietikäinen, M., Matějka, P., Černocký, J., Poh, N., Kittler, J., Larcher, A., Lévy, C., Matrouf, D., Bonastre, J., Tresadern, P., and Cootes, T. (2012) Bi-Modal Person Recognition on a Mobile Phone: using mobile phone data. In: *International Conference on Multimedia and Expo Workshops,*
13. Li, S. Z. and. Jain, A.K. (2005) *Handbook of Face Recognition*. USA, Springer Science and Business Media
14. Wayman, J. L., Jain, A.K., Maltoni, D. and Maio, D. (2005) *Biometric Systems – Technology, Design and Performance Evaluation*. USA, Springer Science and Business Media
15. Singh, N., Agrawal, A. and Khan, R.A. (2015) A Critical Review on Automatic Speaker Recognition. *Science Journal of Circuits, Systems and Signal Processing* **4** (2) 14-17
16. Bousquet, P. M., Bonastre, J.F. and Matrouf (2013) Identify the Benefits of the Different Steps in an i-Vector Based Speaker Verification System. Iberoamerican Congress on Pattern Recognition 278-285. Springer, Berlin, Heidelberg.
17. Wu, X., He, R., Sun, Z. and Tan T. (2015) *A Light CNN for Deep Face Representation with Noisy Labels*. arXiv,

18. Sanderson, C. and Paliwal, K. K. (2002) *Information fusion and person verification using speech andf ace information*. IDIAP-RR 02-33, Switzerland, IDIAP Research Institute
19. Parviz, M. and Moin, M. S. (2011) Boosting approach for score level fusion in multimodal biometrics based on AUC maximization. *Journal of information hiding and multimedia signal processing* **2** (1) 51 - 59
20. Kumar, B. S. and Govardhan, A. (2013) Bipartite RankBoost approach for score level fusion of face and palmprint biometrics. *International journal of advanced research in computer science and software engineering* **3** (12) 249 - 255
21. Toh, K. A., Kim, J. and Lee., S. (2008) Maximizing Area under ROC curve for Biometric Scores Fusion. *Pattern Recognition* **41** 3373-3392
22. Cortes, C., Vapnik V. (1995) Support-vector networks. *Machine Learning* **20** (3) 273 - 297
23. Schölkopf, B., and Smola A. J. (2002) *Learning with Kernels*. Cambridge, MA, MIT Press
24. Schapire, R. E. (1990) The Strength of Weak Learnability. *Machine Learning* **5** (2) 197 - 227
25. Freund, Y. S., Robert E. (1997) A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences* **55** (1) 119 - 139
26. Seiffert, C., Khoshgoftaar, T. M., Hulse, J. V. and Napolitano, A. (2010) RUSBoost: A Hybrid Approach to Alleviating Class Imbalance. *Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans* **40** (1) 185 - 197
27. Platt, J. (1999) Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In: *Advances in large margin classifiers,*
28. Cognitec Systems (2013) *FaceVACS-SDK C++ Reference Manual Version 8.9.5.*
29. Nuance Communications (2013) *Nuance Identifier v9: Delivering Solutions for a Safer World*.
30. Phillips, P. J., Grother, P.J., Michaels, R.J., Blackburn, D.M., Tabassi, E., & Bone, J.M. (2003) *Face Recognition Vendor Test 2002: Evaluation Report*. NISTIR 6965,
31. Hsu, C. W., Chang, C.C. and Lin, C.J. (2010) *A Practical Guide to Support Vector Classification*. Taiwan, National Taiwan University

DEFENCE SCIENCE AND TECHNOLOGY GROUP DOCUMENT CONTROL DATA	1. DLM/CAVEAT (OF DOCUMENT)
2. TITLE Face and Voice Fusion for Human Recognition in Non-controlled Environments	3. SECURITY CLASSIFICATION (FOR UNCLASSIFIED LIMITED RELEASE USE (U/L) NEXT TO DOCUMENT CLASSIFICATION) Document (U) Title (U) Abstract (U)
4. AUTHOR(S) Sau Yee Yiu, Dmitri Kamenetsky, Jason Littlefield and Jonathan Willmore	5. CORPORATE AUTHOR Defence Science and Technology Group PO Box 1500 Edinburgh, South Australia, 5111

6a. DST GROUP NUMBER DST-Group-TR-3426	6b. AR NUMBER AR-017-022	6c. TYPE OF REPORT Technical Report	7. DOCUMENT DATE November 2017

8. OBJECTIVE ID	9.TASK NUMBER 07/029	10.TASK SPONSOR SOCOMD/Army

11. MSTC Intelligence Analytics Branch	12. STC Biometrics
13. DOWNGRADING/DELIMITING INSTRUCTIONS	14. RELEASE AUTHORITY Chief, National Security and ISR Division

15. SECONDARY RELEASE STATEMENT OF THIS DOCUMENT

Approved for public release

OVERSEAS ENQUIRIES OUTSIDE STATED LIMITATIONS SHOULD BE REFERRED THROUGH DOCUMENT EXCHANGE, PO BOX 1500, EDINBURGH, SA 5111

16. DELIBERATE ANNOUNCEMENT

No limitations

17. CITATION IN OTHER DOCUMENTS

Yes

18. RESEARCH LIBRARY THESAURUS

Face Recognition, Speaker Recognition, Biometric Fusion, Classification

19. ABSTRACT

The individual performance of biometric technologies such as speaker recognition (SR) and face recognition (FR) has enabled their prolific use in applications worldwide (e.g. FR at airports and SR for access to telephone banking and taxation purposes). However, in challenging environments (e.g. CCTV videos), where the data is of low quality, establishing the identity of non-cooperative individuals is still a difficult task.

This paper documents the verification performance gains possible when fusing low quality face and voice samples at the matching score level. Three normalisation and five classifier-based fusion techniques were evaluated on a real life audio-video dataset ('Mobio'). When compared to the performance of the individual biometrics, all fused results showed a notable improvement.