Password Less Security System Using MultiFactor Biometric Fusion

Girija Chetty, Dat Tran, Dharmendra Sharma and Bala Balachandran
*School of Information Sciences and Engineering
University of Canberra, Australia
Tel: +61-3-6201 2512, Fax: +61-6201 5231
E-mail: girija.chetty@canberra.edu.au

***Abstract*— In this paper, we propose the use of audio and visual biometric features for person authentication instead of traditional single factor passwords and pins for secure access. Experiments performed on different gender specific subsets of data from VidTIMIT and UCBN multimedia databases under clean and noisy conditions show that with multifactor fusion, about 22-30% improvement in performance can be achieved with as compared to single factor-audio only or visual only mode even under noisy acoustic conditions.**

I. INTRODUCTION

Personal safety in public and private buildings has always been a concern, but since September 11, 2001 is receiving more attention. Authentication methods can be grouped into three classes: something you possess as in an ID card, something you know, and something unique about you, such as biometrics. Possessions (e.g. keys) can be easily lost, forged or duplicated. Knowledge can be forgotten as well as shared, stolen, or guessed. The cost of forgotten passwords is high and accounts for 40% - 80% of all the IT help desk calls [8]. Resetting the forgotten or compromised passwords costs as much as 340$/user/year [9]. Biometrics, on the other hand, are inherently secure since they are some unique feature the person physically has. The science of biometrics is an elegant solution to identifying an individual and avoids problems faced by possession-based and knowledge-based security approaches.

The aggregate security level of a system increases as these three authentication approaches are combined in various ways. The least secure approach is based on PINs (Personal Identification Numbers), which can be easily guessed. The system's security level can be improved by adding some possession such as an identification card. An identification card with a single biometric improves security further. Finally, an identification card with multiple biometrics supports a very high security level as illustrated in Figure 1. In this paper we propose an approach for high level security using multiple biometrics.

Recently there has been a lot of interest in multiple biometric authentication systems [5][10]. Each biometric modality has its own limitations, issues, and problems as discussed in the next section. Not all of these can be solved for a single biometric, even with the use of state of the art and the novel algorithms discovered through further research. Hence, a better approach to building a more robust biometric security system involves integrating multiple biometric sensors.

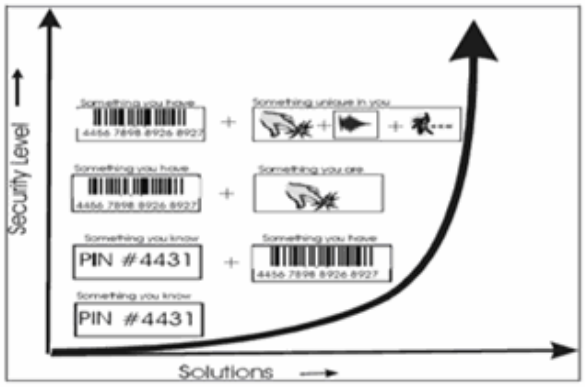


Fig. 1. Solutions to increasing security needs

Currently, the five most common biometric technologies are fingerprinting, iris scanning, hand geometry comparison, face recognition and voice verification. These techniques have significantly varying degrees of accuracy, ease of use, failure to enroll, failure to acquire, and universality. Each technology must perform four basic tasks: biometric acquisition, feature extraction, matching, and decision making.

Face and voice based biometric traits enjoy better user acceptance as compared to other biometrics. There are several reasons for better acceptance of face and voice based biometrics in the society. Human beings have always naturally used face recognition for personal identification purposes. Applications like biometrics, content-based information retrieval, visual surveillance and human computer interaction necessitate successful automation of the recognition task. In automatic face recognition, computer systems are employed to match the test (newly acquired and unknown) face image against a collection of known face

images (training faces) in the database. Although the recognition task seems to be easy and straightforward for people, automated face recognition system becomes challenging and difficult. This is primarily due to the inherent variations in the image acquisition process in terms of image quality, geometry, illumination effects, and occlusion (glasses, facial hair, etc.). These major problems currently limit the accuracy of face recognition.

Face recognition is a very attractive as a biometric because the data is already made public by many of us in the form of a passport or driver's license. Secondly, this biometric can easily be captured by an ordinary camera. Also, the surveillance systems can rely on capturing image without the cooperation of the user. There are many inherent qualities that make it beneficial to automate and improve face recognition.

Voice biometric based verification, like face recognition, is also attractive because of its prevalence in human communication, though its accuracy is currently limited. Voice biometric suffers considerably from variations in the microphone and/or the transmission channel. The performance deteriorates badly as enrollment and use conditions become increasingly mismatched. Background noise can also be a considerable problem. Variations in voice due to illness, emotion or aging are other problems requiring further research.

Due to these limitations, the multifactor biometric fusion does not always lead to synergistic fusion and more often results in what is known as "catastrophic fusion", where the performance of fusion leads to worst performance than the single mode case. This could be due to the focus of most of the research works in proposing novel algorithms and features that perform well under clean audio conditions but lead to a significant performance loss under realistic noisy operating scenarios. Hence the focus of this work is to evaluate the multifactor biometric fusion under clean and noisy acoustic conditions.

To be precise, we propose the fusion of face and voice biometric modalities captured from a video to achieve enhanced security under adverse noise conditions. Experimental results using GMM based speaker models indicate that using multimodal fusion provides significant performance improvement in the level of security. Experiments performed on different gender specific subsets of data from VidTIMIT and UCBN databases under clean and noisy conditions show that with audio-visual fusion, the best EER performance of 5% to 7% can be achieved with multifactor fusion, an improvement of 22-30% as compared to single mode voice only or face only biometric trait.

The paper is organised as follows: The details of audio-visual databases used in the study is described next. The audio-visual fusion process is described Section 3. The Bayesian framework for building Gaussian models is described in section 4. The experimental set up for performing multifactor fusion experiments is described in section 5. The results of the experiments and conclusions derived are described in Sections 6 and 7.

II. AUDIO VISUAL DATABASES

The audio visual data from two different data corpora, VidTIMIT and UCBN was used for evaluating the performance of multimodal fusion features. The VidTIMIT multimodal person authentication database [16], [29], consists of video and corresponding audio recordings of 43 people (19 female and 24 male). The mean duration of each sentence is around 4 seconds, or approximately 100 video frames. A broadcast quality digital video camera in a noisy office environment was used to record the data. The video of each person is stored as a sequence of JPEG images with a resolution of 512□384 pixels with corresponding audio provided as a 16-bit 32-kHz mono PCM file.

The second type of data used is the UCBN database, a free to air broadcast news database. The broadcast news is a continuous source of video sequences, which can be easily obtained or recorded, and has optimal illumination, colour, and sound recording conditions. However, some of the attributes of broadcast news database such as near-frontal images, smaller facial regions, multiple faces and complex backgrounds require an efficient face detection and tracking scheme to be used. The database consists of 20-40 second video clips for anchor persons and newsreaders with frontal/near-frontal shots of 10 different faces (5 female and 5 male). Each video sample is 25 frames per second MPEG2 encoded stream with a resolution of 720 × 576 pixels, with corresponding 16 bit, 48 kHz PCM audio. Figure 2 shows some sample images from the VidTIMIT database (first two rows) and UCBN database (last two rows).

III.

Fig. 2. Audio Visual Databases

IV. MULTIFACTOR FUSION

For multifactor fusion, the audio features and visual features were extracted separately and fused together. The VidTIMIT database was used for experiments described in this paper. The audio signal was divided into frames using a Hamming window of length 20 ms, with a frame overlap of 10 ms to give an audio frame rate, F_A, of 100 Hz. MFCCs of

dimension 12 were extracted from each frame [16]. The audio final audio feature vector consists of these 12 MFCC features, 1 energy component and difference of MFFC features(delta features), thus make it a feature vector of dimension 26. We refer to audio features with f_a notation in the entire paper.

Two types of visual features were extracted, one for the mouth region and one from the entire face. The details of visual feature extraction from mouth region are described in the next section. Similar to audio features, the final visual features comprised of actual visual features concatenated with difference features (delta features) thus capturing both static and dynamic variations in the speaking face.

A. Visual Features

Visual features were extracted from the mouth ROI by automatically segmenting the lip region from rest of the face [17], [18]. We refer to the visual features from the lip region in the entire paper with flip notation. This ROI is segmented manually by locating the two labial corners. A square NpxNp pixel block was extracted as the ROI; where Np = 98 pixels. Due to the lack of head motion of the subjects for the video recordings, manual segmentation was only carried out for every 10th frame, and the ROI coordinates for the intermediate frames were interpolated. Only the gray scale values of the Np x Np were considered. Even though the VidTIMIT and UCBN databases are of high quality, with controlled illumination conditions, the application of histogram equalisation and image demeaning (subtraction of the average pixel intensity values) to the ROI images, were both found to improve the performance of the visual features. Hence these were used to pre-process the images.

Transform based features were used to represent the visual features based on the two dimensional Discrete Cosine Transform (2D-DCT), which was used, because of its high energy compaction and relative superior performance to other image transforms. The 2D-DCT was applied to the pre-processed gray scale pixel blocks. The first 15 coefficients were employed, taken in a "zig-zag" pattern, as illustrated in Figure 3.

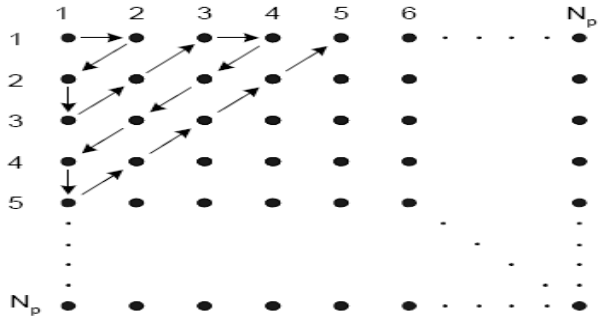


Fig.3. The "zig-zag" manner by which the top 15 DCT features are selected

However, the first coefficient is zero valued (due to the demeaning) and was discarded, leaving 14 static visual features per frame. Calculating the difference of the DCT coefficients across k video frames forms the visual feature vector. This was carried out for two values of k, and via concatenation, this gives a visual feature vector of dimension 30. The values of k employed, depend on the visual feature frame rate. Since the frame rates for visual frames is not normally same as audio frame rates appropriate rate interpolation was done to match the frame rate.

The second type of visual features is popular eigen face features investigated by several works in face recognition area [30]. The eigen face approach is based on principal component analysis, and with controlled illumination, pose and expressions in the face images of the database, it is possible to represent the entire face with 8-10 features, a significant reduction in dimensionality; yielding a satisfactory performance. We refer to eigen face features in the entire paper with ef_{face} notation.

V. Bayesian Framework for Speaker Models

To evaluate the performance of proposed multimodal fusion features, the gender specific GMM speaker models were obtained using text dependent and text independent data subsets of VidTIMIT and UCBN corpora.

A, Gaussian Mixture Speaker Models

For text independent modeling, the speaker does not speak the same utterance during enrolment (training) phase and testing phase, whereas for text dependent modeling, the speaker uses the same utterance during enrolment phase and testing phase. The speaker models were obtaining by building a large gender-specific universal background models (UBMs) first, and then adapting the UBMs to speaker models similar to the approach described [26]. The advantage of using UBM is that the impostor likelihood is now speaker independent. Moreover, it was found by several researchers [26],[29] that instead of constructing the client models directly from the training data (using EM algorithm), lower error rates can be obtained (on a larger database) when the client models are generated by adapting the UBM using a form of MAP adaptation [16],[26],[29]. A full description of MAP adaptation is out of the scope of this paper (the reader is encouraged to refer to [16][29]. The update equations are summarized as follows:

Given UBM parameters, where N_G is the number of Gaussian mixtures, and a set of training feature vectors for a specific client, , the estimated mixture weights ($\widehat{m}_k$), means ($\hat{\mu}_k$), and covariances ($\hat{\Sigma}_K$) are found by iterative expectation

maximization algorithm [Conrad]. The final parameters are obtained as shown below in Eqn. 1:

$$c_k = [\alpha \hat{c}_k + (1-\alpha)\dot{c}_k]\gamma$$
$$\vec{\mu}_k = \alpha \hat{\vec{\mu}}_k + (1-\alpha)\dot{\vec{\mu}}_k$$
$$\Sigma_k = [\alpha(\hat{\Sigma}_k + \hat{\vec{\mu}}_k \hat{\vec{\mu}}_k^T) + (1-\alpha)(\dot{\Sigma}_k + \dot{\vec{\mu}}_k \dot{\vec{\mu}}_k^T))] - \vec{\mu}_k \vec{\mu}_k^T \quad (1)$$

Where α is a scale factor to make sure all weights sum to one. γ is a data dependent adaptation coefficient (L_k and r , a fixed relevance factor is described in more detail in [16],[29]). It must be noted that UBM mixture components will only be adapted if there is a sufficient correspondence with client training data. Thus to prevent the final client models not being specific enough (leading to poor performance), the UBM must adequately represent the general client population). This was the reason behind using separate male and female UBMs for our experiments here.

VI. PERFORMANCE EVALUATION

Since the verification system is inherently a two-class decision task, it follows that the system can make two types of errors. The first type of error is a False Acceptance Error (FA), where an impostor is accepted. The second error is a False Rejection (FR), where a true claimant is rejected. Thus the performance is measured in terms of False Acceptance Rate (FAR %) and False Reject Rate (FRR %), can be defined as (Eqn. 2):

$$FAR\ \% = \frac{I_A}{I_T} \times 100\ \%$$
$$FRR\ \% = \frac{C_R}{C_T} \times 100\ \% \quad (2)$$

where I_A is the number of impostors classified as true claimants, I_T is the total number of impostor classification tests, C_R is the number of true claimants classified as impostors, and C_T is the total number of true claimant classification tests.

Since the errors are related, minimizing the FAR % increases the FRR % (and vice versa). The trade-off between FAR % and FRR % is adjusted using the threshold t, an experimentally determined speaker independent global threshold from the training/enrolment data. The trade-off between FAR % and FRR % can be graphically represented by a Receiver Operating Characteristics (ROC) plot or a Detection Error Trade-off (DET) plot [29]. The ROC plot is on a linear scale, while the DET plot is on a normal-deviate logarithmic scale. For DET plots, the FRR% is plotted as a function of FAR %. To quantify the performance into a single number, Equal Error Rate (EER) is often used [29]. Here the system is configured with a threshold, set to an operating point when FAR % = FRR %.

It must be noted that the threshold can also be adjusted to obtain desired performance on test data (data unseen by the system up to this point). Such a threshold is known as the *aposteriori* threshold. However, if the threshold is fixed before finding the performance, the threshold is known as the apriori threshold [38]. The *apriori* threshold can be found via experimental means using training/enrolment or evaluation data (data which has also been unseen by the system up to this point, but is separate from test data).

Logically, the *apriori* threshold is more realistic. However, it is often difficult to find a reliable *apriori* threshold [16], [29]. The test section of a database is often divided into two sets: evaluation data and test data. If the evaluation data is not representative of the test data, then the *apriori* threshold will achieve significantly different results on evaluation and test data. Moreover, such a database division reduces the number of verification tests, thus decreasing the statistical significance of the results. For these reasons, many researchers prefer to use the *apostereriori* and interpret the performance obtained as the expected performance. For all the single-mode experiments in this paper, we have used *apostereriori* threshold (from test set) so that a comparison could be made with some of the existing approaches.

A. *Late Fusion vs. Feature Fusion*

The two main problems concerning multifactor fusion are when and how the fusion should take place. As reviewed in [10] on audio-visual approaches, three common levels to carry out the fusion, include; the early (feature-level), the middle-level, and the late (score-level) [14],[15],[17] fusion. Feature fusion and late fusion only for the proposed multifactor fusion approach is examined in all experiments in this paper.

- Feature Fusion

The features extracted from the audio and visual modalities (implicit and explicit lip motion features) were concatenated, and then used for the training and testing of an audio-visual GMM speaker model. Eqn. (3) and

$$O_n^{\{AV\}} = [o_n^{\{A\}}, o_n^{\{V\}}],\ 1 \le n \le N_A \quad (3)$$

- Late Fusion

Late fusion requires two independent classifiers to be trained, one classifier for each modality. Due to independent processing of modalities it does not preserve the acoustic-labial dynamics and correlation properties. However, some advantages of late fusion include; the ability to account for modality reliabilities, small feature vector dimensions, and ease of adding other modalities to the system. For late fusion, the two scores are weighted to account for the reliability of the modalities. The two scores may be integrated via addition or multiplication. Eqn. (4) shows the use of weights for the

case of additive integration, where β_A and $\beta_V (= 1 - \beta_A)$ are the weights placed on the audio and visual scores respectively. The audio and visual features need not be rate-interpolated for late fusion as they are processed independently by separate classifiers. Prior to late fusion the audio and visual scores are normalized, so that each set of client and impostor scores fall into the range [0, 1].

$$P(O_{AV} \mid S_i) = \beta_A \times P(O_A \mid S_i) + (1 - \beta_A) \times P(O_v \mid S_i) \quad (4)$$

VII. Experimental Setup

In this section, different experiments conducted to evaluate the performance of proposed multimodal fusion features for the biometric system performance are described. For performance evaluation, different subsets of data from VidTIMIT and UCBN were used. The gender-specific universal background models (UBMs) were developed using training data from two sessions, Session 1 and Session 2 of the VidTIMIT corpus and for testing stage session 3 was used. Due to the type of data available (test session sentences different from training session sentences), only text independent experiments could be performed with VidTIMIT database. This gave 1536 (2*8*24*4) seconds of training data for the male UBM and 576(2*8*19*4) seconds of training data for the female UBM. The GMM topology with 10 Gaussian mixtures was used for all the experiments. The number of Gaussian mixtures was determined empirically to give the best performance. For UCBN database, similar gender-specific universal background models (UBMs) were obtained using training data from text dependent subset 1 and text independent subset 3 & 4. Ten sessions of the male and female speaking face data from these subsets were used for training and 5 sessions for testing.

For testing if the numbers of subjects are R, the impostors are generated by the leave-one-out scheme, with each subject being the impostor for the remaining R-1 subjects. For all the experiments, the threshold was set using test data. For male only subset for VidTIMIT database, there were 48 client trials (24 male speakers x 2 test utterances in session 3), and 1104 impostor trials (24 male speakers x 2 test utterances in session 3 x 23 impostors/client), and for the female VidTIMIT subset, there were 38 client trials (19 male speakers x 2 test utterances in session 3), and 684 impostor trials (19 male speakers x 2 test utterances in session 3 x 18 impostors/client). For male only subset for UCBN database, there were 25 client trials (5 male speakers x 5 test utterances in each subset), and 100 impostor trials (5 male speakers x 5 test utterances x 4 impostors/client), and for the female UCBN subset, there were similar number of client and impostor trials as in male subset as there we used 5 male and female speakers from different subsets. Different sets of experiments were conducted to evaluate the performance of the multimodal fusion features in terms of DET curves and equal error rates (EER).

To examine the effects of background acoustic noise, additive white Gaussian noise was applied to the clean audio data at SNRs ranging from 48dB to –12dB in steps of 6dB. The audio/audio-visual models were trained using the original clean audio data and tested on audio data with the various SNR levels. The best fusion weight (β_A) values for the late fusion were determined empirically by exhaustive search for each audio SNR test level. This was achieved by testing β_A values ranging from 0 to 1 in steps of 0.01. The β_A value was chosen such that the speaker verification performance at the given test set SNR test level was maximised.

VIII. EXPerimental Results

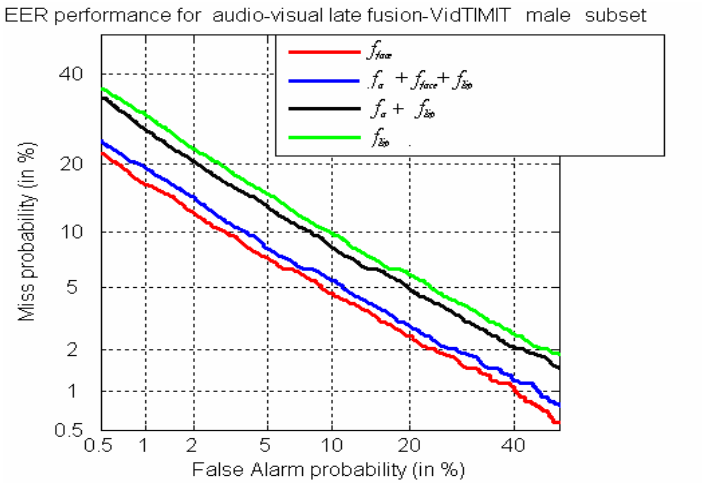
The experimental results are presented with the following organisation. First we compare the results of single mode and multifactor fusion for clean audio case. Next we compare the performance of the multifactor fusion features across various audio SNR levels.

A. *Evaluation of Multifactor Fusion Features for Clean Audio Case*

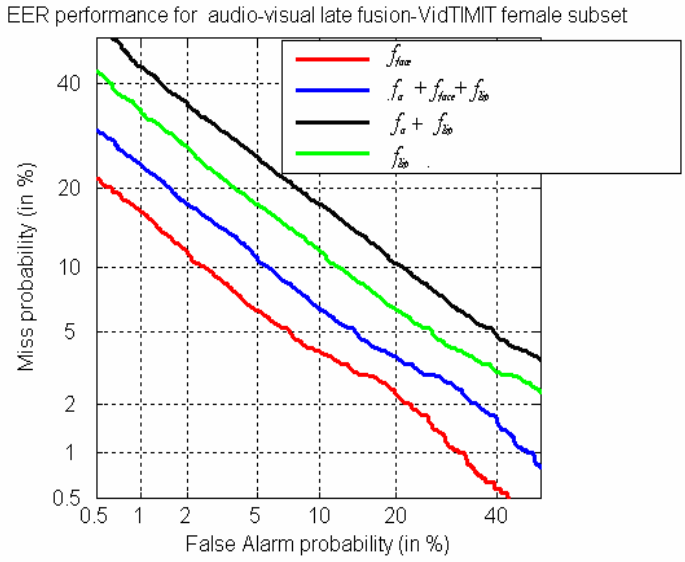
Table 1 and DET curves in Figure 3 to 6 show the EER performance with late fusion of audio and visual features. As can be seen in Table 1, and the DET curves in Figure 3,4,5 and 6, the best EER performance of 6.53 % is achieved with late fusion of audio, lip and face features ($f_a+f_{lip}+f_{face}$) for the VidTIMIT male subset.

Table 1: EER performance for late fusion of audio, face and lip features

Feature Set	VidTIMIT male subset EER (%)	VidTIMIT female subset EER (%)	UCBN male subset EER (%)	UCBN female subset EER (%)
f_a	4.38	5.01	5.35	5.46
f_{face}	5.1	5.5	6.72	6.73
f_{lip}	9.22	12.3	12.86	13.1
$f_a + f_{face} + f_{lip}$	6.53	7.5	8.53	8.64
$f_a + f_{face}$	8.15	9.1	10.87	11.15
$f_a + f_{lip}$	8.6	10.99	13.77	13.52



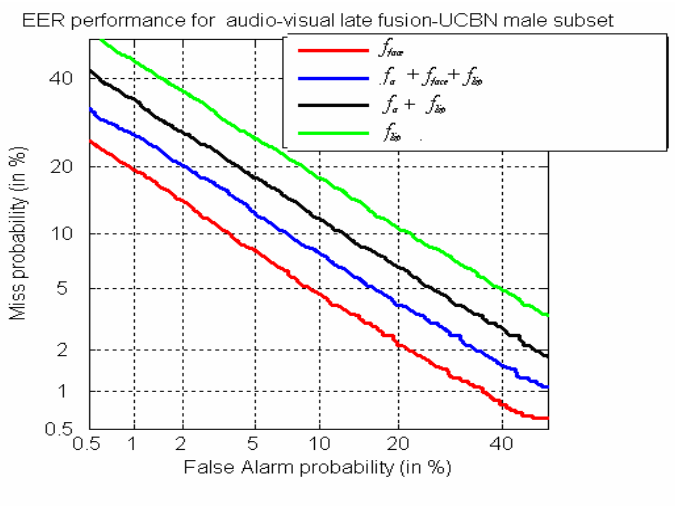
Fig,3. DET curves for evaluating audio visual late fusion: VidTIMIT male dataset



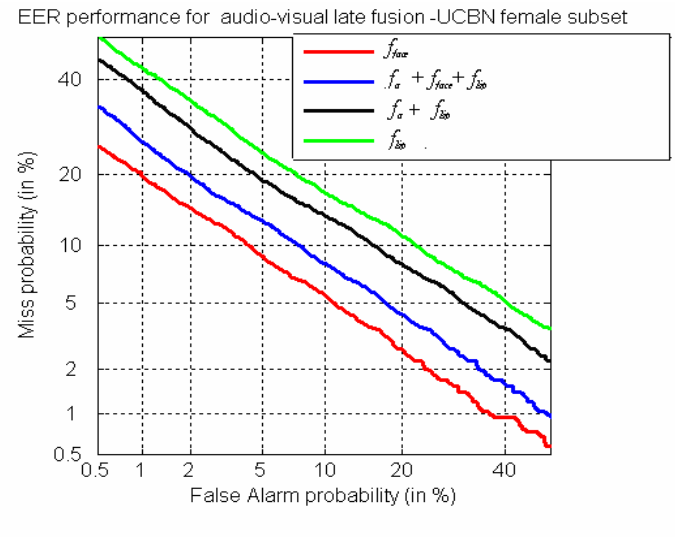
Fig,4. DET curves for evaluating audio visual late fusion: VidTIMIT female dataset

B. *Evaluation of Multimodal Fusion Features for Noisy Audio Case*

The EER performance shown in Table 1 and Table 2 is for the clean audio case. As can be seen in the first row of Table 1 and 2, the EER performance with clean audio is better than audio-visual fusion case. The "clean" audio refers to the original VidTIMIT/UCBN acoustic speech data, prior to the application of any degradation. It can be expected that the audio modality would perform well for the clean speech. To examine the effects of background acoustic noise, additive white Gaussian noise was applied to the clean audio at SNRs ranging from 48 dB to -20 dB in steps of 6 dB.



Fig,5. DET curves for evaluating audio visual late fusion: UCBN male dataset



Fig,6. DET curves for evaluating audio visual late fusion:UCBN female dataset

Table 2: EER performance for late fusion of audio, face and lip features

Feature Set	VidTIMIT male subset EER (%)	VidTIMIT female subset EER (%)	UCBN male subset EER (%)	UCBN female subset EER (%)
f_a	4.38	5.01	5.35	5.46
f_{face}	5.0	5.5	6.5	6.6
f_{lip}	9.15	12.3	12.8	13
$f_a + f_{face} + f_{lip}$	6.46	7.5	8.48	8.57
$f_a + f_{face}$	8.05	9.1	15.97	16.08
$f_a + f_{lip}$	8.45	10.72	13.75	13.35

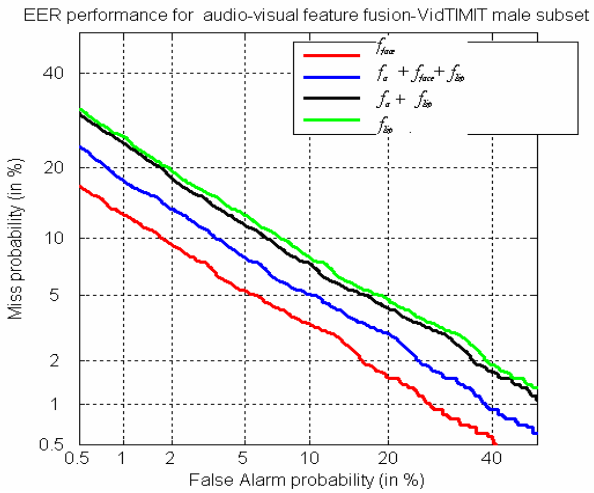


Fig. 7. DET curves for evaluating audio visual feature fusion male VidTIMIT dataset

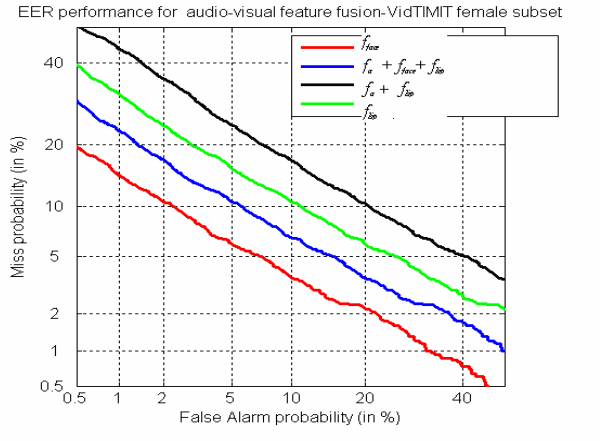


Fig. 8. DET curves for evaluating audio visual feature fusion female VidTIMIT dataset

The EER performance achieved with feature level fusion of audio and implicit and explicit lip motion features is shown in Table 2 and the DET curves in Figure 7 to 10. It can be observed that the feature level fusion leads to similar performance improvement as late level fusion, though it is cited in some audio visual fusion literature [21,22], that feature level fusion leads to performance loss as compared to late fusion.

Moreover, the fusion of audio, face and lip features performs better as compared to the fusion of audio and face, audio and lip features for VidTIMIT male subset. This could be due to the ability of fusion of all three features to model the person identity better.

The audio/audio-visual GMM models were trained using the original clean audio data and tested on audio data with various

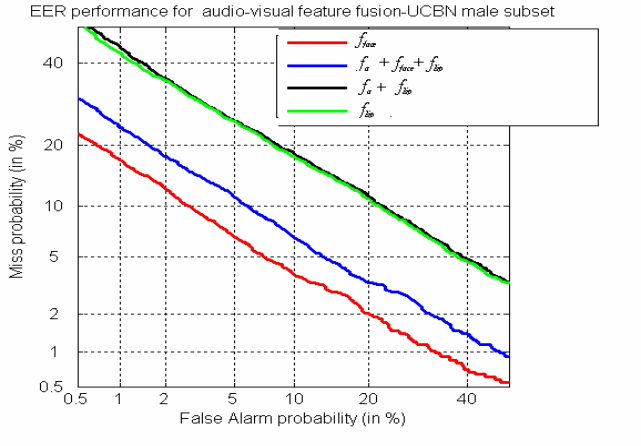


Fig. 9. DET curves for evaluating audio visual feature fusion male UCBN dataset

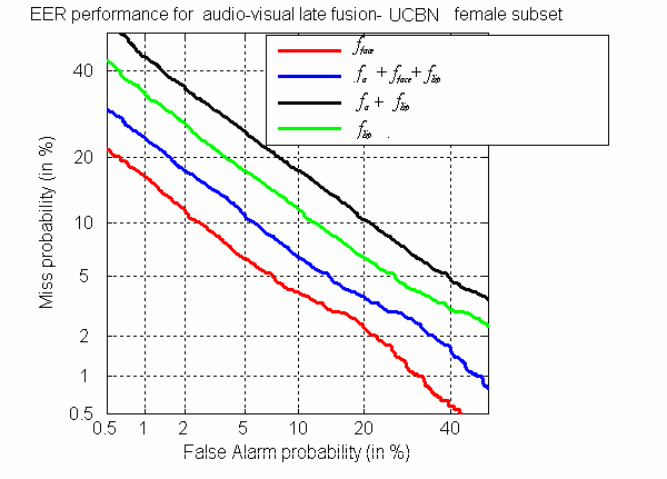


Fig. 10. DET curves for evaluating audio visual feature fusion male UCBN dataset

SNR levels. Late fusion was performed as in Eqn. (4). The best values were determined empirically by exhaustive search for each audio SNR test level. This was achieved by testing values ranging from 0 to 1 in steps of 0.01. The values was chosen such that the identity verification performance at a given test set SNR test level was maximized (i.e. there was prior knowledge of the correct decisions).

At higher noise levels, however, the audio only performance deteriorates significantly; whereas, the visual features (lip and face) features, and the fusion of audio and visual features perform better than audio performance. The results for the noisy acoustic data are now presented.

Figures 11 and 12 compare the EER performance of late-fusion and feature-fusion approach for VidTIMIT and UCBN datasets across various audio SNR levels. For male VidTIMIT subset, the audio accuracy roll off with respect to test-set SNR level is high, with an audio EER of 4.38% at 48dB which drops to around 100% EER at –6dB. This steep audio performance roll off with respect to SNR is due to the

mismatched audio testing conditions, i.e., training on noise-free audio and testing on audio of a lower SNR. It is expected that the roll off would be less steep if matched testing was employed, i.e., training and testing using audio of the same SNR.

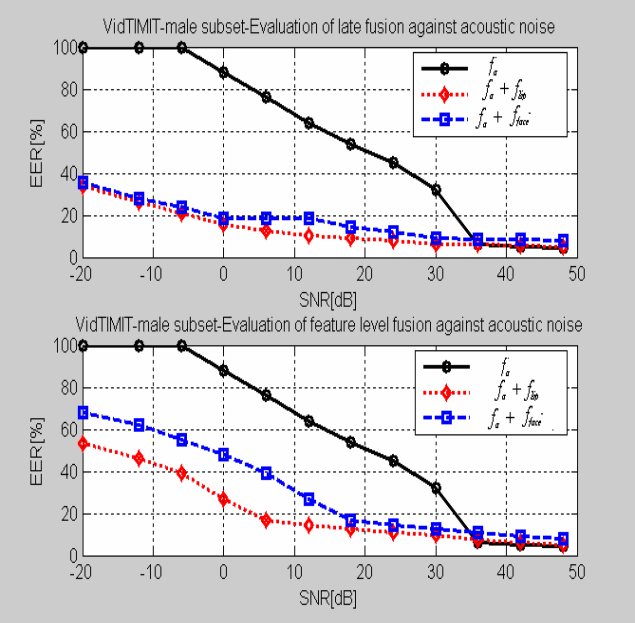


Fig.11. Evaluation of late fusion and feature fusion under noisy acoustic test environment for VidTIMIT male subset

However, multifactor fusion features perform better in noisy acoustic conditions, and in general, the EER performance of late fusion features is better than feature level fusion with both face and lip region visual features. As can be expected, the visual only features are unaffected by noisy test conditions with an EER of 6.8% for features and 9.3% for features even at low audio SNRs.

With audio-visual multifactor fusion, both face and lip region fusion features yielded a similar performance. For male VidTIMIT subset, until 30 dB SNR, the late fusion of (fa + flip) and (fa + fface), allows a synergestic fusion with EERs less than visual only EERs. However for feature fusion, the system is more sensitive to noisy test conditions and the fusion is synergestic for less than 42 dB SNRs as compared to 30dB SNR threshold for late fusion. Also, at low SNRs, the feature fusion leads to a higher performance roll off, whereas for late fusion, the system continues to be more robust even at low SNR levels. Similar performance roll off can be observed for female data subset of UCBN, with the EER performance for all modes (audio, visual, and audio-visual fusion) relatively poorer, due to lesser training data available for female UCBN subset as compared to male VidTIMIT subset.

There may be several reasons why the feature fusion, leads to catastrophic fusion irrespective of whether audio-face and audio-lip features are used. This could be due to adverse affect a corrupted audio feature vector can have on the audio-visual fusion vector at very low audio SNRs. Another reason for poor feature fusion performance as compared to late fusion may be lack of sufficient training data for training the GMM speaker models trained with concatenated audio-visual feature vectors with larger dimensions. This could also be due to the GMM topology used for modelling the speakers. Alternate topologies based on HMMs may allow building of better speaker models, with higher number of HMM states/mixes, and this would have boosted the performance. However, this would require more training data as compared to the training data available from VidTIMIT and UCBN databases used for experiments here.

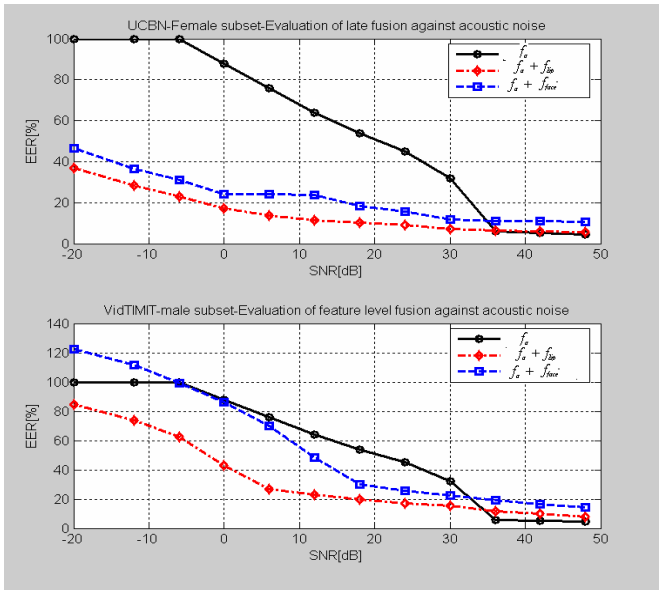


Fig.12. Evaluation of late fusion and feature fusion under noisy acoustic test environment for UCBN female subset

IX. CONCLUSIONS

The empirical results presented in this paper on multifactor fusion based on audio, face and lip region features are quite promising, particularly showing that the addition of the visual modality not only improve the performance at low SNR test levels but also enhances the performance for clean audio, resulting in the performance with higher robustness to audio noise. It was also shown that multifactor fusion of all three (audio, face and lip) features better than audio-face and audio-lip features, and the late fusion approach of acoustic and visual features leads to better performance. This is due to better representation of person' specific information with audio and two types of visual features. In conclusion, the results show that the multifactor fusion of biometric features, makes the identity verification system robust against acoustic noise degradations.

X. REFERENCES

[1] John D. Woodward, Jr., “Biometrics: Facing Up to Terrorism”, The Biometrics Consortium Conference 2002, Arlington Virginia, February, 2000.

[2] Stephen King, “Personal Identification Pilot Study,” The Biometrics Consortium Conference 2002, Arlington, Virginia, February, 2002.

[3] Lin Hong and Anil Jain, “Integrating Faces and Fingerprints for Personal Identification”, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 20, No. 12, Dec., 1998, pp. 1295 - 1307.

[4] Salil Prabhakar and Anil Jain, “Decision-level Fusion in Fingerprint Verification”, Pattern Recognition, vol. 35, 2002, pp. 861-874.

[5] Sharath Panikanti , Ruud M. Bolle , Anil Jain, “Biometrics : The Future of Identification”, IEEE Computer, Vol. 33, No. 2, February 2000.

[6] A.K. Jain, S. Prabhakar, S. Chen, “Combining multiple matchers for a high security fingerprint verification system”, Pattern Recognition Letters 20 (11-13) (1999) 1371-1379.

[7] Forrester Research, Inc, http://www.forrester.com, 2001.

[8] Gartner Group, http://www.gartner.com , 2001 .

[9] Tony Mansfield, Gavin Kelly, David Chandler, and Jan Kane, Biometric Product Testing Final Report, Computing, National Physical Laboratory, Crown Copyright, UK, March, 2001.

[10] Robert W. Frischholz , Ulrich Deickmann, “ BioID: A Multimodal Biometric Identification System” , IEEE Computer, Vol. 33, No. 2, February 2000.

[11] L.Hong, A.K. Jain, S. Panikanti, “Can multibiometrics improve perfomance?”,Proceedings AutoID’99, Summit, NJ, October 1999, pp. 59-64.

[12] L.Osadciw, P.K.Varshney, and K. Veeramachaneni, “ Improving Personal Identification Accuracy Using Multisensor Fusion for Building Access Control Applications”, Proceedings of the Fifth International Conference on Information Fusion, July 2002, Annapolis, Maryland.

[13] Pramod K. Varshney , Distributed Detection and Data Fusion, Springer, New York,1997.

[14] A. Ross and A. Jain, "Information fusion in biometrics," Pattern Recognition Letters, vol. 24, pp. 2115-2125, 2003/9 2003.

[15]. Pankanti, R. M. Bolle, and A. Jain, "Biometrics:The future of identification," Computer, vol. 33, pp. 46-49 2000.

[16]C. Sanderson and K. K. Paliwal, "Multi-modal person verification system based on face profiles and speech," Proceedings of the Fifth International Symposium on Signal Processing and Its Applications, (ISSPA). Brisbane, vol. 2, pp. 947-950, Aug. 1999.

[17] R. Brunelli and D. Falavigna, "Person Identification Using Multiple Cues," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 17, pp. 955-966, Oct. 1995.

[18] C. C. Chibelushi, F. Deravi, and J. S. D. Mason, "A Review of Speech-Based Bimodal Recognition," IEEE Transactions on Multimedia, vol. 4, pp. 23-35, Mar 2002.

[19] K. A. Toh, W. Y. Yau, and X. Jiang, "A Reduced Multivariate Polynomial Model for Multimodal Biometrics and Classifiers Fusion," IEEE Transactions on Circuits and Systems For Video Technology, vol. 14, pp. 224-233, Feb. 2004.

[20] P. J. Phillips, et al., "Face Recognition Vendor Test 2002," IEEE International Workshop on Analysis and Modeling of Faces and Gestures, (AMFG), Nice, France, pp. 44, Oct. 2003.

[21] J. P. Campbell, "Speaker Recognition: A Tutorial," Proceedings of the IEEE, vol. 85, pp. 1437-1462, Sept. 1997.

[22] M. Turk and A. Pentland, "Eigenfaces for Recognition," Journal of Cognitive Neuroscience, vol. 3, pp. 71-86 1991.

[23] J. M. Naik, "Speaker Verification: A Tutorial," Communications Magazine, IEEE, vol. 28 pp. 42 - 48, Jan. 1990.

[24] R. O. Duda, P. E. Hart, and D. G. Stork, "Pattern Classification": Wiley-Interscience, 2001.

[25] S. Theodoridis and K. Koutroumbas, "Pattern Recognition": Academic Press, 1999.

[26] D. A. Reynolds and R. C. Rose, "Robust Text-Independent Speaker Identification sing Gaussian Mixture Speaker Models," IEEE Tran. on Speech and Audio Processing, vol. 3 pp. 72-83, Jan. 1995.

[27] R. Brunelli and T. Poggio, "Face Recognition: Features versus Templates," Pattern Analysis and Machine Intelligence, IEEE Transactions on, vol. 15, pp. 1042-1052 1993.

[28] A. V. Nefian and I. Hayes, M. H., "Hidden Markov models for face recognition," Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Seattle, Washington, pp. 2721–2724, May 1998.

[29] C. Sanderson and K. K. Paliwal, "Fast features for face authentication under illumination direction changes," Pattern Recognition Letters, vol. 24, pp. 2409-2419, 2003/10 2003.

[30] M. A. Turk and A. P. Pentland, "Face recognition using eigenfaces," Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition, (CVPR), pp. 586-591 1991.