Biometric Liveness Checking Using Multimodal Fuzzy Fusion

Girija Chetty, *Member IEEE*

***Abstract*—In this paper we propose a novel fusion protocol based on fuzzy fusion of face and voice features for checking liveness in secure identity authentication systems based on face and voice biometrics. Liveness checking can detect fraudulent impostor attacks on the security systems, and ensure that biometric cues are acquired from a live person who is actually present at the time of capture for authenticating the identity. The proposed fuzzy fusion of audio visual features is based on mutual dependency models which extract the spatio-temporal correlation between face and voice dynamics during speech production, Performance evaluation in terms of DET (Detector Error Tradeoff) curves and EERs (Equal Error Rates) on publicly available audiovisual speech databases show a significant improvement in performance of proposed fuzzy fusion of face-voice features based on mutual dependency models over conventional fusion techniques.**

I. Introduction

MOST of the commercial biometric security systems currently deployed are based on modeling the identity of a person based on unimodal biometric information, i.e. fingerprint, face, or voice features. Also, authentication schemes for many current interactive civilian human computer interaction applications are based on speech based voice features, which achieve significantly lower performance for operating environments with low signal-to-noise ratios (SNR). Use of both visual and audio information can lead to better robustness, as they can provide complementary secondary clues that can help in the analysis of the primary biometric signals [1]. For instance, it is well known that deaf people can learn how to lip read. The joint analysis of co-occurring acoustic and visual speech signals during speech production can improve the robustness of automatic recognition systems [2, 3].

There is a significant body of work on use of joint face-voice information for improving the performance of identity authentication systems. However, most of these state-of-the-art approaches are based on independently processing the voice and face information and then fusing the scores – score fusion [4,5,6]. A major weakness of these systems is that they do not take into account the dominant and non-dominant joint relationship between audio and visual cues, leading to fraudulent impostor attacks on systems in real world unsupervised scenarios, such as for on-line web based authentication applications. Due to the technological advances in computer graphics and animation technologies, it has become easier for impostors to artificially synthesize the biometric data, crafting a replay attack. Such fraudulent replay attacks leave the systems vulnerable to spoofing by recording the voice of the target in advance and replaying it in front of the microphone, or simply placing a still picture of the target's face in front of the camera. Such problems can be approached with liveness check methods, which ensure that biometric cues are acquired from a live person who is actually present at the time of capture for authenticating the identity. With the diffusion of Internet based authentication systems for day-to-day civilian scenarios happening at a astronomical pace [7], it is high time to think about use of user-friendly non-intrusive biometrics for identity authentication, and face and voice biometrics rate high on this aspect. However, face and voice rate high also in terms of vulnerability to tampering and fraudulent replay, and use of effective countermeasures such as liveness checking needs to be incorporated in biometric security systems for better diffusion of biometric authentication technologies in day-to-day real world operating scenarios. Though there is some work in finger print based liveness detection techniques [8,9], there is hardly any work on approaches on liveness checks based on user-friendly biometric identifiers (face and voice)

A significant progress however, has been made in independent processing of face only or voice only based authentication approaches [1,2,3,4,5,6].These approach do not take into consideration an inherent coupling that exists between jointly occurring dominant and non-dominant biometric identifiers. Some preliminary approaches (such as the one described in [7, 8] address liveness checking problem by jointly modeling the acoustic and visual speech features for testing liveness. They involve the fusion of acoustic, appearance and shape based visual features from lip region for jointly modeling the co-occurring face-voice dynamics in speaking face video sequences. However, they are heuristic and ad hoc approaches that work in ideal laboratory environments, and do not perform well in real world unsupervised operating scenarios.

In this paper we propose a novel fusion technique based on fuzzy fusion for joint analysis of acoustic and visual speech features for modeling liveness information. The rest of the paper is organized as follows. Section 2 describes the motivation for using mutual dependency features, and the proposed liveness check approach based on fusion is described in Section 3. Section 4 details the data corpora used and the experimental evaluation of the proposed fusion approach, with Section 5 summarizing the conclusions drawn from this work and plans for further research.

Manuscript received February 7, 2010. The. author is with the Faculty of Information Sciences and Engineering, University of Canberra, Australia (phone: 61-2-62012512; fax: 61-2-62015231; e-mail: girija.chetty@canberra.edu.au .

978-1-4244-8126-2/10/$26.00 ©2010 IEEE

II. MOTIVATION FOR JOINT FEATURE ANALYSIS

The motivation to use joint feature analysis technique for modeling the spatio-temporal relationship between co-occurring face and voice signals during speech production is based on the following two observations: The first observation is in relation to any video event, for example a speaking face video, where the content usually consists of the co-occurring audio and the visual elements. Both the elements carry their contribution to the highest level semantics, and the presence of one has usually a "priming" effect on the other: for example, when hearing a dog barking we expect the image of a dog, seeing a talking face we expect the presence of her voice, images of a waterfall usually brings the sound of running water etc. A series of psychological experiments on the cross-modal influences [9, 10] have proved the importance of synergistic fusion of the multiple modalities in the human perception system. A typical example of this kind is the well-known McGurk effect [9]. Several independent studies by cognitive psychologists suggest that the type of multi-sensory interaction between acoustic and orafacial articulators occurring in the McGurk effect involves both the early and late stages of integration processing [9,10]. It is likely that a human brain uses a hybrid form of fusion that depends on the availability and quality of different sensory cues.

Yet, in audiovisual identity verification systems, the analysis is usually performed separately on different modalities, and the results are brought together using different types of heuristic and ad hoc fusion methods. However, in this process of separation of modalities or the technique used for extraction of features, we lose valuable joint cross-modal information about the whole event or the object we are trying to analyze and detect. There is an inherent association between the two modalities (although complex) and the analysis should take advantage of the synchronized appearance of the relationship between the audio and the visual signal. The second observation relates to different types of fusion techniques used for joint processing of audiovisual speech signals. The late-fusion strategy, which comprises decision or the score fusion, is effective especially in case the contributing modalities are uncorrelated and thus the resulting partial decisions are statistically independent. Feature level fusion techniques, on the other hand, can be favoured (only) if a couple of modalities are highly correlated.

However, jointly occurring face and voice dynamics during speech production is neither highly correlated (mutually dependent) nor loosely correlated nor totally independent (mutually independent). A complex and nonlinear spatiotemporal coupling consisting of highly coupled, loosely coupled and mutually independent components may exist between co-occurring acoustic and visual speech signals [11, 12]. The compelling and extensive findings by authors in [11] validate such complex relationship between external face movements, tongue movements, and speech acoustics when tested for consonant vowel (CV) syllables and sentences spoken by male and female talkers with different visual intelligibility ratings. They proved that the there is a higher correlation between speech and lip motion for C/a/ syllables than for C/i/ and C/u/ syllables. Further, the degree of correlation differs across different places of articulation, where lingual places have higher correlation than bilabial and glottal places. Also, mutual coupling can vary from talker to talker; depending on the gender of the talker, vowel context, place of articulation, voicing, and manner of articulation and the size of the face. Their findings also suggest that male speakers show higher correlations than female speakers. Further, the authors in [12] also validate the complex, spatiotemporal and non-linear nature of the coupling between the vocal-tract and the facial articulators during speech production, governed by human physiology and language-specific phonetics. They also state that most likely connection between the tongue and the face is indirectly by way of the jaw. Other than the biomechanical coupling, another source of coupling is the control strategy between the tongue and cheeks. For example, when the vocal tract is shortened the tongue does not get retracted.

Due to such a complex nonlinear spatiotemporal coupling between speech and lip motion, this could form a good candidate for detecting liveness. Modeling the speaking faces by capturing this information can make the biometric authentication systems less vulnerable to spoof and fraudulent replay attacks, as it would be almost impossible to spoof a system which can accurately distinguish the artificially manufactured or synthesized speaking face video sequences from the live video sequences. We propose an approach based on joint audio-visual features and subsequent fusion based on fuzzy logic to address this problem. Next two Sections briefly describe the two main aspects of the proposed approach.

III. JOINT FEATURE EXTRACTION

Joint feature extraction based on Canonical Correlation Analysis (CCA) was used for extracting the audio and visual signals from the speaking face video sequences.. The CCA was first proposed by Hotelling [13], and is a method of determining a linear space where the correlations between two sets of variables are maximized. This approach has been successfully applied to sets of variables that are manifestations of a set of hidden variables, examples of this are fMRI and image retrieval[14]. The audio-visual speaking face modeling is a similar candidate, since the motions of articulators and the speech produced are fundamentally linked. However, CCA is derived as a linear process and this limitation becomes apparent in the cases where the underlying relationship is non-linear [15], such as the complex nonlinearity in correlation relationship between the speech and lip-motion during speech production. To circumvent this linearity restriction, we have used a "kernel trick", which allows replacing an inner product by a projection of the data into a higher dimensional space, and performing CCA in this realized dual representation [15].

We perform a kernel Canonical Correlation Analysis (kCCA) on Mel Frequency Cepstral Coefficients (MFCC) voice features and the lip motion features extracted from a biological inspired optical flow algorithm called Multi Channel Gradient Model (MCGM).

The MCGM is a neuro-physiological and psycho-physically inspired unified motion algorithm [15]. In MGCM approach, the behavior of V1/V2 cells is modeled by MGCM functions and the ratio of temporal and spatial gradients is computed to establish local velocity estimates. From one sequence of lip region images it is possible to derive two sets of visual information from MCGM, initially a sequential series of frames are anlaysed by MCGM algorithm, calculating the relative motion between successive frames. Additionally, a current frame of data is processed against a fixed open mouth frame, calculating the absolute motions of the mouth. MCGM processing results in a matrices of equal size to the input frames, each containing speed and angular information for a given pixel. Applying (linear) Principal Component Analysis (PCA) produces a linear space onto which the motions can be mapped, reducing the dimensionality of the visual features.

Mel-Frequency Cepstral Coefficients (MFCC) are classical acoustic speech features used in automatic speech processing [16]. They are state-of-the-art features in many applications, including automatic speech recognition and speaker verification systems. For obtaining a MFFC feature vector, the voice signal is transformed into the frequency domain via windowed Fast Fourier Transform and then mapped on to the Mel scale, a human perceptual scale of frequency [16]. A (logarithmically spaced) filter bank is constructed over this Mel frequency spectrum, and from this the logarithm of the power spectrum is determined. A discrete time cosine transform is performed over the power spectrum and the MFCCs are calculated. Most of the information about human voice from speech can be captured by retaining 10-12 most significant MFCC features, the first-order time-derivatives(delta features), the pitch and the signal energy.

To account for the lack of synchronization between speech features and lip motion features, rate interpolation can be done by up sampling the MCGM features to obtain the synchronized MCGM-MFCC features. Once the acoustic MFCC features and MCGM lip motion features are obtained, kCCA is implemented by first mapping them onto the kernel space using polynomial kernels and then performing CCA. Since, the kCCA involves, implementing CCA in a higher dimensional nonlinear space, it has the capability to capture and track the nonlinear correlations between different features. Parameter tuning for kCCA can be performed offline on an independent data set.

For extracting the mutually independent components of the audio and visual signals, another powerful statistical technique called independent component analysis (ICA) is performed, which treats the observed variables as a mixture of independent sources. Two different approaches can be used for Independent Component Analysis, ICA1 and ICA2 [17, 18]. In ICA1, the basis images are independent, whereas in ICA2 the mixing coefficients are independent. We utilize the ICA2 approach, where each pixel for lip images are considered as a mixture of independent coefficients. If X is a data matrix incorporating the measured variables, then it can be split as: X = AS where A is the mixing matrix and S contains the independent coefficients. The columns of A form a basis for the database and the columns of S provide ICA-features for the corresponding lip images residing in the columns of the data matrix X.

For each pixel, all x and y coordinates of a lip image are concatenated to a single vector. Its dimensionality is then reduced by applying PCA to the training set of x-y co-ordinate vectors. Each face is then represented by the first K PCA coefficients. The columns of the data matrix X for the ICA analysis are constituted of PCA coefficient vectors. Then, the Fast ICA algorithm described by [17, 18] is applied to obtain the basis A and the independent coefficients S. Next section describes the .proposed fuzzy fusion technique used to combine various features.

IV. FUZZY FUSION TECHNIQUE

First, we derive the algorithm for performing the fuzzy fusion using multiple features described in the previous Section. Let us denoted the projection of audio and lip features in each of the closely coupled (kCCA), and mutually independent (ICA) subspaces as f_{kCCA} and f_{ICA}. We also include the projection of visual information in the PCA subspace as Eigenlip features f_{PCA} as the static spatial information in face images contains identity specific information as well.

According to Medasani et al [19] and Keller et al. [20], the most generic way of performing the fuzzy fusion is to normalize, fuzzify, compute the fuzzy integral (fusion) and defuzzify. Figure 1 shown below describes the main steps for fuzzy fusion technique and the scheme is described below. For simplicity, the fusion between two feature sets i.e. kCCA features and MCGM features is described.

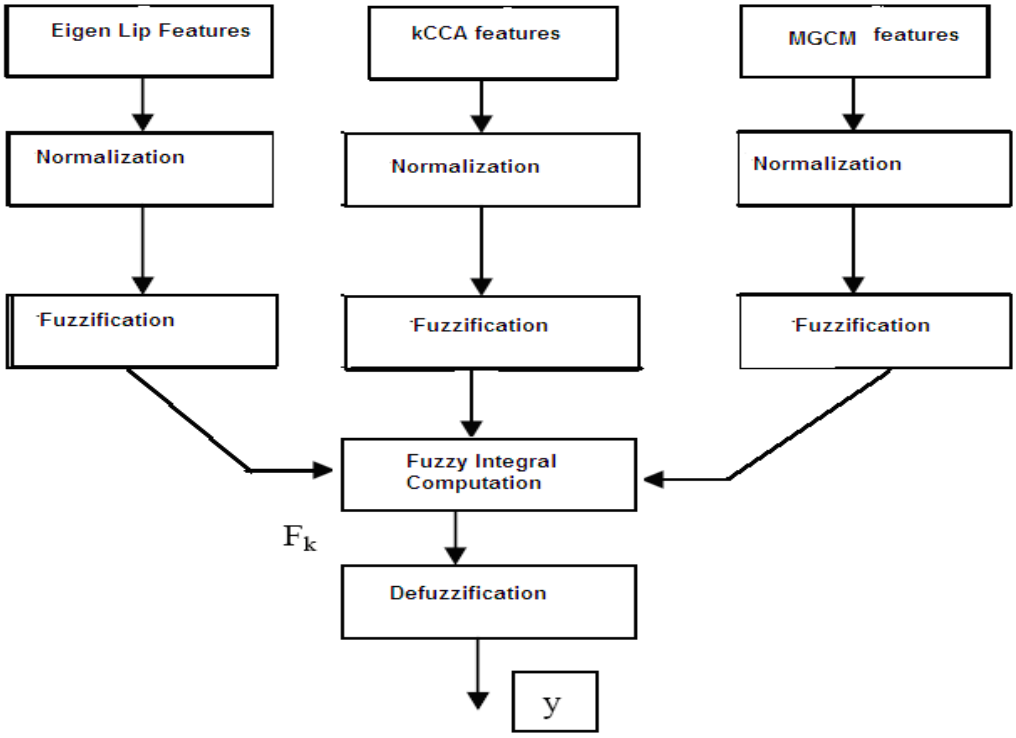


Fig. 1. Block Schematic of the Fuzzy Fusion Scheme

Step 1. We first compute Eigen lip, kCCA and MGCM features from the video frames.

Step 2. The kCCA, MGCM and Eigen lip features were normalized prior to fusion. The normalized vector v of an original vector is defined as

$$v = \frac{\Omega}{\sqrt{(\Omega^T \Omega)}} \tag{1}$$

Step 3. The purpose of fuzzification is to map input vector v from each modality to values from 0 to 1, representing evidence that the object satisfies the class hypothesis *kC*. The generation of membership function is very important [19] [20]. In this paper, we propose a histogram-based method for generating the membership function.

Let x be the distance between input object and its class, and $h(x)$ be the histogram of x, which provides information regarding the distribution of distance. Membership function $u(x)$ can be constructed as follows

$$u(x) = \int_{x}^{+\infty} h(x)dx \tag{2}$$

From Eq. 2, we construct membership function $u(x)$ for each feature vector. Let

$$\varepsilon_k = \|v - v_k\|,$$

where v_k is the vector describing the k th class. The fuzzification result S_k is computed as

$$S_k = \mu(\varepsilon_k) \tag{3}$$

Step 4. Fuzzy integral considers the objective evidence supplied by each source (called the *h*-function) and the expected worth of each source (via a fuzzy measure) [19][20]. Let x_1 represent the kCCA features, x_2 represent the MGCM features, and x_3 represent the Eigenlip features. The fuzzy density value $g^i = g\{x_i\}$ is determined via statistical measurements on errors rates of the single modality x_i. Thus the output of fuzzy integral F_k *can* be expressed as:

$$F_k = \begin{cases} \max(\min(S_{kV}, g^1), S_{kA}) \\ \max(\min(S_{kA}, g^1), S_{kV}) \end{cases} \begin{matrix} S_{kV} > S_{kA} \\ else \end{matrix} \tag{4}$$

where S_{kV} *is* the fuzzification result of *kCCA* features, and S_{kA} *is* the fuzzification result of *MGCM* features.

Step 5. We classify the kCCA and MGCM features into a specific class if the fuzzy integral F_k had the output of fuzzy integral:

$$y = \arg\max_{k} F_k \tag{5}$$

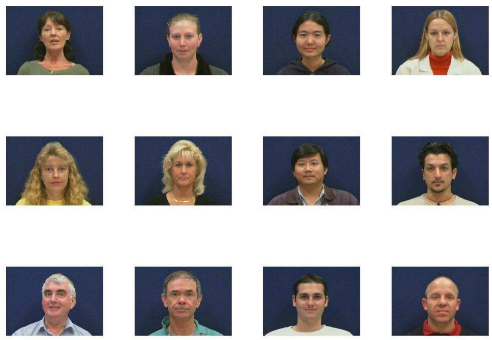
V. Experimental results

Experiemental evalution was done with two audio-visual speaking face video corpora VidTIMIT [21] and DaFEx [22,23], showing a significant improvement in liveness check performance due to the use of fuzzy fusion technique and multiple correlation features (kCCA, MCGM, ICA and PCA). Figure 2 shows some images from the two corpora used for experiemental work. The details of the two corpora are given in [21] , [22] and [23]. For a comparative evaluation, experiments were also performed with convention late fusion technique, a deterministic fusion approach based on equal fusion weights for each set of features.

The experiments involve a training phase and a test phase. The training phase involves building of statistical models based on Gausian mixtures from the training data sets. The testing stage for the liveness check scenario is different from the traditional testing in biometric identity verification scenarios. Here the impostor data is artificially synthesised replay attack test data emulating fraudulent attacks. Two different types of replay attacks were tested, the static replay attacks with still photo and audio, and the dynamic replay attacks, where artificial speaking face sequences are synthesised from still photo, few key frames from the video sequences, and lip-synched with pre-recorded speech signals.

A 10-mixture Gaussian mixture model λ of a client's audiovisual feature vectors was built in the training phase, reflecting the probability densities for the combined phonemes and visemes (lip shapes) in the audiovisual feature space. In the testing phase, the clients' live test recordings were first evaluated against the client's model λ by determining the log likelihoods log p(X|λ) of the time sequences X of audiovisual feature vectors under the usual assumption of statistical independence of successive feature vectors.

For testing static replay attacks, a number of "fake" or synthetic recordings were constructed by combining the sequence of audio feature vectors from each test utterance with ONE visual feature vector chosen from the sequence of visual feature vectors and keeping that visual feature vector constant throughout the utterance. Such a synthetic sequence represents an attack on the authentication system, carried out by replaying an audio recording of a client's utterance while presenting a still photograph to the camera. Four such fake audiovisual sequences were constructed from different still frames of each client test recording. Log-likelihoods log p(X'|λ) were computed for the fake sequences X' of audiovisual feature vectors against the client model λ. In order to obtain suitable thresholds to distinguish live recordings from fake recordings, detection error trade-off (DET) curves and equal error rates (EER) were determined. For testing dynamic replay attacks artificially synthesized speaking face video sequences were used instead of using the actually recorded video sequences in the data corpora.



(a) VidTIMIT corpus images

(b) DaFeX corpus images

Fig. 2. Face Images from VidTIMIT and DaFex Corpus

Since the liveness checking is a two-class decision task, the system can make two types of errors. The first type of error is a False Acceptance Error (FA), where an impostor (fraudulent replay attacker) is accepted. The second error is a False Rejection (FR), where a true claimant (genuine client) is rejected. Thus, the performance is measured in terms of False Acceptance Rate (FAR) and False Reject Rate (FRR), as defined as (Eqn. 6) :

$$FAR\ \% = \frac{I_A}{I_T} \times 100\ \%$$
$$FRR\ \% = \frac{C_R}{C_T} \times 100\ \% \tag{6}$$

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. The implications of this is minimizing the

FAR increases the FRR and vice versa, since the errors are related. The trade-off between FAR and FRR is adjusted using the threshold θ, an experimentally determined person--independent global threshold from the training or 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. The ROC plot is on a linear scale, while the DET plot is on a normal-deviate logarithmic scale. For DET plot, the FRR is plotted as a function of FAR. To quantify the performance into a single number, the Equal Error Rate (EER) is often used. Here the system is configured with a threshold, set to an operating point when FAR % = FRR %.

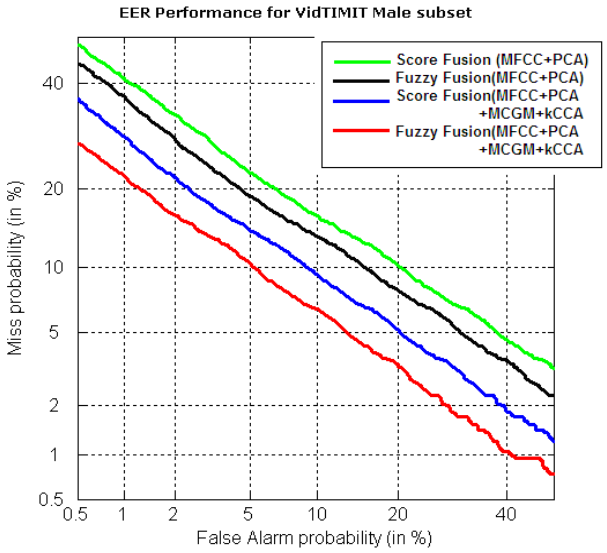
It must be noted that the threshold θ can also be adjusted to obtain a desired performance on test data (data unseen by the system up to this point). Such a threshold is known as the aposteriori threshold. However, if the threshold is fixed before finding the performance, the threshold is known as the apriori threshold. 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.

Practically, the apriori threshold is more realistic. However, it is often difficult to find a reliable apriori threshold. 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 aposteriori and interpret the performance obtained as the expected performance.

Different sets of experiments were conducted to evaluate the performance of the audio-visual correlation features based on proposed fuzzy fusion of correlation features (kCCA) and motion (MCGM) features). The performance evaluation for different features in terms of DET curves and EERs is shown n Table 1 to Table 4, and Figure 3 and Figure 4. As can be seen from Figure 3, Figure 4 and Table 1, the results for VidTIMIT data set are quite promising for fuzzy fusion of kCCA correlation features and MCGM motion features when combined with features in ICA and PCA space as compared to conventional fusion based on equal weights for each type of features. As can be seen in Table 2, 3 and 4, similar improvement can be seen for other data sets as well.

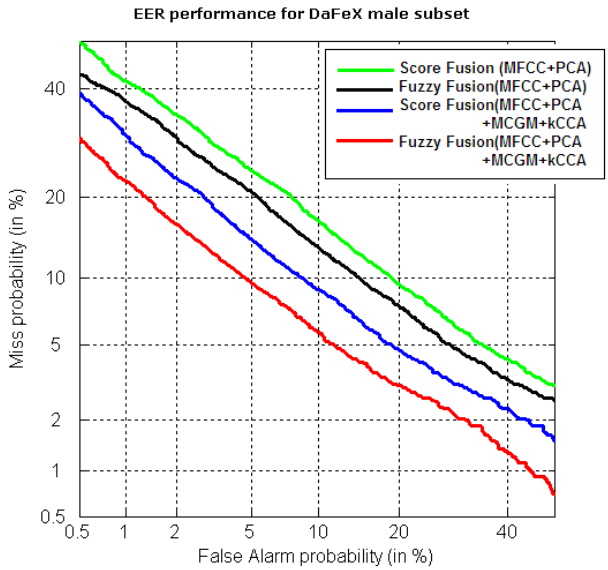
Further, the DET curves show not only the importance of fuzzy fusion technique but also the role of the joint correlation and motion features. This is because the use of PCA and ICA features on their own does not have the capability to capture the hidden correlation relationship between co-occurring signals. However, when combined with correlation and motion features extracted by kCCA and MCGM technique and combined with fuzzy fusion approach, they result in significant improvement in verification error rates.

Moreover, the error rates achieved comparatively for DaFeX corpora show the robustness of the approach for more challenging operating scenario (face data with expression variations). As expected, the performance achieved for this data is lower as compared to VidTIMIT. However, DaFex data represents more challenging and realistic operating scenario as compared to VidTIMIT, representing a laboratory setting suitable for benchmarking purposes.



(a)

Figure 3: DET curves for audio visual features based on fuzzy fusion of mutual dependency features for VidTIMIT data set



(b)

Figure 4: DET curves for audio visual features based on fuzzy fusion of mutual dependency features for DaFeX dataset

TABLE I
ERROR RATES (EERS) FOR VIDTIMIT MALE DATASET

Audio Visual Features	Score Fusion % EER	Fuzzy Fusion % EER
$f_{mfcc} + f_{eigLip}$	16.8 %	16.2%
$f_{mfcc} + f_{eigLip} + f_{kCCA}$	17.2 %	14.7%
$f_{mfcc} + f_{eigLip} + f_{kCCA} + f_{MGCM}$	13.03%	11.68%
$f_{mfcc} + f_{eigLip} + f_{kCCA} + f_{MGCM} + f_{ICA}$	10.26%	**8.06%**

TABLE 2
ERROR RATES (EERS) FOR VIDTIMIT FEMALE DATASET

Audio Visual Features	Score Fusion % EER	Fuzzy Fusion % EER
$f_{mfcc} + f_{eigLip}$	16.88 %	16.2%
$f_{mfcc} + f_{eigLip} + f_{kCCA}$	17.87 %	15.18%
$f_{mfcc} + f_{eigLip} + f_{kCCA} + f_{MGCM}$	14.12%	11.86%
$f_{mfcc} + f_{eigLip} + f_{kCCA} + f_{MGCM} + f_{ICA}$	10.26%	**8.85%**

TABLE 3
ERROR RATES (EERS) FOR DAFEX MALE DATASET

Audio Visual Features	Score Fusion % EER	Fuzzy Fusion % EER
$f_{mfcc} + f_{eigLip}$	15.7 %	16.64%
$f_{mfcc} + f_{eigLip} + f_{kCCA}$	15.9%	14.81%
$f_{mfcc} + f_{eigLip} + f_{kCCA} + f_{MGCM}$	13.12%	11.79%
$f_{mfcc} + f_{eigLip} + f_{kCCA} + f_{MGCM} + f_{ICA}$	10.46%	**9.23%**

TABLE 4
ERROR RATES (EERS) FOR DAFEX FEMALE DATASET

Audio Visual Features	Score Fusion % EER	Fuzzy Fusion % EER
$f_{mfcc} + f_{eigLip}$	15.7 %	15.7 %
$f_{mfcc} + f_{eigLip} + f_{kCCA}$	15.54 %	15.28%
$f_{mfcc} + f_{eigLip} + f_{kCCA} + f_{MGCM}$	14.4%	11.17%
$f_{mfcc} + f_{eigLip} + f_{kCCA} + f_{MGCM} + f_{ICA}$	10.46%	**9.31%**

VI. CONCLUSIONS

In this paper we proposed a novel fuzzy fusion technique for liveness checking in biometric security systems based on combining multimodal features extracted using kernel kCCA and MCGM features, which model the close coupling between audio and visual signals during speech production, and combine with PCA and ICA features which model independent and loose coupling. Performance evaluation in terms of DET curves and EERs on VidTIMIT and DaFeX corpora, showed a significant reduction in verification error rates.

REFERENCES

[1] Gerasimos Potamianos, Chalapathy Neti, Juergen Luettin, and Iain Matthews. Audio-Visual Automatic Speech Recognition: An Overview. Issues in Visual and Audio-Visual Speech Processing, 2004.

[2] Xiaoxing Liu, Luhong Liang, Yibao Zhaa, Xiaobo Pi, and Ara V. Nefian. Audio-Visual Continuous Speech Recognition using a Coupled Hidden Markov Model. In Proc. International Conference on Spoken Language Processing, 2002.

[3] Sabri Gurbuz, Zekeriya Tufekci, Tufekci Patterson, and John N. Gowdy. Multi-Stream Product Modal Audio-Visual Integration Strategy for Robust Adaptive Speech Recognition. In Proc. IEEE International Conference on Acoustics, Speech and Signal Processing, Orlando, 2002.

[4] Claude C. Chibelushi, Farzin Deravi, and John S.D.Mason. A Review of Speech-Based Bimodal Recognition. IEEE Transactions on Multimedia, 4(1):23–37, 2002.

[5] Hao Pan, Zhi-Pei Liang, and Thomas S. Huang. A New Approach to Integrate Audio and Visual Features of Speech. In Proc. IEEE International Conference on Multimedia and Expo., pages 1093 – 1096, 2000.

[6] U.V. Chaudhari, G.N. Ramaswamy, G. Potamianos, and C. Neti. Information Fusion and Decision Cascading for Audio-Visual Speaker Recognition Based on Time- Varying Stream Reliability Prediction. In IEEE International Conference on Multimedia Expo., volume III, pages 9 – 12, Baltimore, USA, July 2003.

[7] Chetty G., and Wagner M,, Robust face-voice based speaker identity verification using multilevel fusion, Image and Vision Computing, Volume 26, Issue 9, 1 September 2008, Pages 1249-1260.

[8] R.Goecke and J.B. Millar. Statistical Analysis of the Relationship between Audio and Video Speech Parameters for Australian English. In J.L. Schwartz, F. Berthommier, M.A. Cathiard, and D. Sodoyer (eds.), Proceedings of the ISCA Tutorial and Research Workshop on Auditory-Visual Speech Processing AVSP 2003, pages 133-138, St. Jorioz, France, 4 - 7 September 2003.

[9] S. Molholm, et al., "Multisensory Auditory-visual Interactions During Early Sensory Processing in Humans: a high-density electrical mapping study," Cognitive Brain Research, vol. 14, pp. 115-128, June 2002.

[10] J. MacDonald, & H. McGurk, "Visual influences on speech perception process". Perception and Psychophysics, 24, 253-257, 1978.

[11] J.Jiang, A. Alwan, P.A.Keating, E.T. Auer Jr., L. E. Bernstein, "On the Relationship between Face Movements, Tongue Movements, and Speech Acoustics," EURASIP Journal on Applied Signal Processing 2002:11, 1174–1188.

[12] H. C. Yehia, T. Kuratate, and E. Vatikiotis-Bateson, "Using speech acoustics to drive facial motion," in Proc. the 14th International Congress of Phonetic Sciences, pp. 631–634, San Francisco, Calif, USA, 1999.

[13] D. R. Hardoon, S. Szedmak and J. Shawe-Taylor, "Canonical Correlation Analysis: An Overview with Application to Learning Methods", in Neural Computation Volume 16, Number 12 2004, Pages 2639–2664.

[14] H. Hotelling. "Relations between two sets of variates." Biometrika, 28:321 377, 1936.

[15] P.W. McOwan, and A. Johnston, "The algorithms of natural vision: The Multi-channel Gradient Model". Proc. IEE/IEEE Genetic Algorithms in Engineering Systems. Sept' 95.

[16] S. Chauhan and P. Wang and C.S. Lim and V. Anantharaman "A computer-aided MFCC-based HMM system for automatic auscultation" Comput. Biol. Med., Vol. 38, No. 2, 2008, Pages 221–233.

[17] K. W. Bowyer, K. Chang, and P. Flynn, "A survey of approaches and challenges in 3D and multi-modal 3D + 2D face recognition," Computer Vision and Image Understanding, vol. 101, no. 1, pp. 1–15, 2006.

[18] A. Srivastava, X. Liu, and C. Hesher, "Face recognition using optimal linear components of range images," Image and Vision Computing, vol. 24, no. 3, pp. 291–299, 2006.

[19] Medasani, S., Kim, J., Krishnapuram, R.: An overview of Membership Function Generation Techniques for Patter Recognition. International Journal of Approximate Reasoning 19 (1998) 391–417.

[20] Keller, J. M., Osborn, J: Training the Fuzzy Integral. International Journal of Approximate Reasoning 15 (1996) 1–24.

[21] C. Sanderson. Biometric Person Recognition: Face, Speech and Fusion. VDM-Verlag, 2008. ISBN 978-3-639-02769-3.

[22] Battocchi, A.; Pianesi, F.. 2004. DaFEx: Un Database di Espressioni Facciali Dinamiche. In Proceedings of the SLI-GSCP Workshop, Padova (Italy) 30 Novembre - 1 Dicembre 2004.

[23] Mana N., Cosi P., Tisato G., Cavicchio F., Magno E. and Pianesi F., An Italian Database of Emotional Speech and Facial Expressions, In Proceedings of "Workshop on Emotion: Corpora for Research on Emotion and Affect", in association with "5th International Conference on Language, Resources and Evaluation (LREC2006), Genoa, Italy, 24-25-26 May 2006.