1

Multimodal Fusion for Robust Identity Authentication: Role of Liveness Checks

Girija Chetty and Emdad Hossain
Faculty of Information Sciences and Engineering, University of Canberra, Australia

1. Introduction

Most of the current biometric identity authentication systems currently deployed are based on modeling the identity of a person based on unimodal information, i.e. face, voice, or fingerprint features. Also, many current interactive civilian remote 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). For a long time, use of acoustic information alone has been a great success for several automatic speech processing applications such as automatic speech transcription or speaker authentication, while face identification systems based visual information alone from faces also proved to be of equally successful. However, in adverse operating environments, performance of either of these systems could be suboptimal. 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 (Potamianos et al (2004)). The joint analysis of acoustic and visual speech can improve the robustness of automatic speech recognition systems (Liu et al (2002), Gurbuz et al (2002).

There have been several systems proposed on use of joint face-voice information for improving the performance of current identity authentication systems. However, most of these state-of-the-art authentication approaches are based on independently processing the voice and face information and then fusing the scores – the score fusion (Chibelushi et al (2002), Pan et al (2000), Chaudari et. al.(2003)). A major weakness of these systems is that they do not take into account fraudulent replay attack scenarios into consideration, leaving them 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. This problem can be addressed with liveness verification, which ensures 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 at a astronomical pace (Chetty and Wagner (2008)), it is high time to think about the vulnerability of traditional biometric authentication approaches and consider inclusion of liveness checks for next generation biometric systems. Though there is some work in finger print based liveness checking techniques (Goecke and Millar (2003), Molhom et al (2002)), there is hardly any work in liveness checks based on user-

friendly biometric identifiers (face and voice), which enjoy more acceptability for civilian Internet based applications requiring person identity authentication.
A significant progress however, has been made in independent processing of face only or voice only based authentication approaches (Chibelushi et al (2002), Pan et al (2000), Chaudari et. al.(2003)), in which until now, inherent coupling between jointly occurring primary biometric identifiers were not taken into consideration. Some preliminary approaches such as the ones described in (Chetty and Wagner (2008), Goecke and Millar (2003)), address liveness checking problem by using the traditional acoustic and visual speech features for testing liveness. Both these approaches, neither considered an inherent coupling between speech and orafacial articulators (lips, jaw and chin) during speech production, nor used a solid pattern recogntion based evaluation framework for the validating the performance of co-inertia features.
In this Chapter we propose a novel approach for extraction of audio-visual correlation features based on cross-modal association models, and formulate a hybrid fusion framework for modelling liveness information in the identity authentication approach. Further, we develop a sound evaluation approach based on Bayesian framework for assessing the vulnerability of system at different levels of replay attack complexity. The rest of the Chapter is organized as follows. Section 2 describes the motivation for using the proposed approach, and the details the cross-modal association models are described in Section 3. Section 4 describes the hybrid fusion approach for combining the correlation features with loosely couple and mutually independent face-speech components. The data corpora used and the experimental setup for evaluation of the proposed features is described in Section 5. The experimental results, evaluating proposed correlation features and hybrid fusion technique is discussed in Section 6. Finally, Section 7 summarises the conclusions drawn from this work and plans for further research.

2. Motivation for cross modal association models

The motivation to use cross-modal association models 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: 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 bring the sound of running water etc. A series of psychological experiments on the cross-modal influences (Molhom et al (2002), MacDonald and McGurk (1978)) 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 (MacDonald and McGurk (1978)). 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 (MacDonald and McGurk (1978)). 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 speech and speaker verification systems, the analysis is usually performed separately on different modalities, and the results are brought together using different fusion methods. However, in this process of separation of modalities, we lose

valuable cross-modal information about the whole event or the object we are trying to analyse and detect. There is an inherent association between the two modalities and the analysis should take advantage of the synchronised 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 in speaking face video sequences, 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 in speaking face video sequences (Jiang et al(2002), Yehia et al (1999)). The compelling and extensive findings by authors in Jiang et al (2002), 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 Yehia et al (1999), 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 be an ideal candidate for detecting and verifying liveness, and modelling 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. Next section briefly describes the proposed cross modal association models based on cross-modal association models.

3. Cross-modal association models

In this section we describe the details of extracting audio-visual features based on cross-modal association models, which capture the nonlinear correlation components between the audio and lip modalities during speech production. This section is organised as follows: The details of proposed audio-visual correlation features based on different cross modal association techniques: Latent Semantic Analysis (LSA) technique, Cross-modal Factor Analysis (CFA) and Canonical Correlation Analysis (CCA) technique is described next.

3.1 Latent Semantic Analysis (LSA)

Latent semantic analysis (LSA) is used as a powerful tool in text information retrieval to discover underlying semantic relationships between different textual units e.g. keywords and paragraphs (Li et al(2003), Li et al(2001)). It is possible to detect the semantic correlation between visual faces and their associated speech based on the LSA technique. The method consists of three major steps: the construction of a joint multimodal feature space, the normalization, the singular value decomposition (SVD), and the semantic association measurement.

Given n visual features and m audio features at each of the *t* video frames, the joint feature space can be expressed as:

$$X = [V_1, \ldots, V_i, \ldots, V_n, A_1, \ldots, A_i, \ldots A_m] \tag{1}$$

where

$$V_i = (v_i(1), v_i(2), \ldots, v_i(t))^T \tag{2}$$

and

$$A_i = (a_i(1), a_i(2), \ldots, a_i(t))^T \tag{3}$$

Various visual and audio features can have quite different variations. Normalization of each feature in the joint space according to its maximum elements (or certain other statistical measurements) is thus needed and can be expressed as:

$$\hat{X}_{ij} = \frac{X_{kl}}{\max\limits_{k,l}(abs(X_{kl}))} \qquad \forall k,l \tag{4}$$

After normalisation, all elements in the normalised matrix $\hat{X}$ have values between –1 and 1. SVD can then be performed as follows:

$$\hat{X} = S \cdot V \cdot D^T \tag{5}$$

where S and D are matrices composed of left and right singular vectors and V is the diagonal matrix of singular values in descending order.

Keeping only the first k singular vectors in S and D, we can derive an optimal approximation of with reduced feature dimensions, where the semantic correlation information between visual and audio features is mostly preserved. Traditional Pearson correlation or mutual information calculation (Li et al (2003), Hershey and Movellan (1999), Fisher et al(2000)) can then be used to effectively identify and measure semantic associations between different modalities. Experiments in Li et al(2003), have shown the effectiveness of LSA and its advantages over the direct use of traditional correlation calculation.

The above optimization of $\hat{X}$ in the least square sense can be expressed as:

$$\hat{X} \cong \tilde{X} = \tilde{S}.\tilde{V}.\tilde{D} \tag{6}$$

Where $\tilde{S}, \tilde{V}, and \tilde{D}$ consist of the first *k* vectors in *S*, *V*, and *D*, respectively.

The selection of an appropriate value for *k* is still an open issue in the literature. In general, *k* has to be large enough to keep most of the semantic structures. Eqn. 6 is not applicable for applications using off-line training since the optimization has to be performed on the fly directly based on the input data. However, due to the orthogonal property of singular vectors, we can rewrite Eqn. 6 in a new form as follows:

$$\hat{X} \cong \tilde{X} = \tilde{S} \cdot V \cdot \tilde{D}^T \tag{7}$$

Now we only need the $\tilde{D}$ matrix in the calculation, which can be trained in advance using ground truth data. This derived new form is important for those applications that need off-line trained SVD results.

3.2 Cross Modal Factor Analysis (CMA)

LSA does not distinguish features from different modalities in the joint space. The optimal solution based on the overall distribution, which LSA models, may not best represent the semantic relationships between the features of different modalities, since distribution patterns among features from the same modality will also greatly impact the results of the LSA.

A solution to the above problem is to treat the features from different modalities as two separate subsets and focus only on the semantic patterns between these two subsets. Under the linear correlation model, the problem now is to find the optimal transformations that can best represent or identify the coupled patterns between the features of the two different subsets. We adopt the following optimization criterion to obtain the optimal transformations:

Given two mean-centred matrices X and Y, which consist of row-by-row coupled samples from two subsets of features, we want orthogonal transformation matrices A and B that can minimise the expression:

$$\|XA - YB\|_F^2 \tag{8}$$

where

$A^T A = I$ and $B^T B = I$.

$\|M\|_F$ denotes the Frobenius norm of the matrix M and can be expressed as:

$$\|M\|_F = \left(\sum_i \sum_j |m_{ij}|^2 \right)^{1/2} \tag{9}$$

In other words, *A* and *B* define two orthogonal transformation spaces where coupled data in *X* and *Y* can be projected as close to each other as possible.

Since we have:

$$\|XA - YB\|_F^2 = trace\left((XA - XB).(YA - YB)^T\right)$$

$$= trace\left(XAA^T X^T + YBB^T Y^T - XAB^T Y^T - YBA^T X^T\right) \tag{10}$$

$$= trace\left((XX^T) + trace(YY^T) - 2 \cdot trace(XAB^T Y^T)\right)$$

where the trace of a matrix is defined to be the sum of the diagonal elements. We can easily see from above that matrices A and B which maximise *trace* (XAB^TY^T) will minimise (10). It can be shown (Li et al(2003)), that such matrices are given by:

$$\begin{cases} A = S_{xy} \\ B = D_{xy} \end{cases} \quad \text{where } X^T Y = S_{xy} \cdot V_{xy} \cdot D_{xy} \tag{11}$$

With the optimal transformation matrices A and B, we can calculate the transformed version of X and Y as follows:

$$\begin{cases} \tilde{X} = X \cdot A \\ \tilde{Y} = Y \cdot B \end{cases} \tag{12}$$

Corresponding vectors in $\tilde{X}$ and $\tilde{Y}$ are thus optimised to represent the coupled relationships between the two feature subsets without being affected by distribution patterns within each subset. Traditional Pearson correlation or mutual information calculation (Li et al (2003), Hershey and Movellan(1999), Fisher et al(2000)) can then be performed on the first and most important k corresponding vectors in $\tilde{X}$ and $\tilde{Y}$, which similar to those in LSA preserve the principal coupled patterns in much lower dimensions.
In addition to feature dimension reduction, feature selection capability is another advantage of CFA. The weights in A and B automatically reflect the significance of individual features, clearly demonstrating the great feature selection capability of CFA, which makes it a promising tool for different multimedia applications including audiovisual speaker identity verification.

3.3 Canonical Correlation Analysis (CCA)

Following the development of the previous section, we can adopt a different optimization criterion: Instead of minimizing the projected distance, we attempt to find transformation matrices A and B that maximise the correlation between X_A and Y_B. This can be described more specifically using the following mathematical formulations:
Given two mean centered matrices X and Y as defined in the previous section, we seek matrices A and B such that

$$correlation(XA, XB) = correlation(\tilde{X}, \tilde{Y}) = diag(\lambda_1, \cdots \lambda_i, \cdots, \lambda_l) \tag{13}$$

where $\tilde{X} = Y \cdot B$, and $1 \geq \lambda_1 \geq, \cdots, \lambda_i, \cdots, \geq \lambda_l \geq 0$. λ_i represents the largest possible correlation between the ith translated features in $\tilde{X}$ and $\tilde{Y}$. A statistical method called canonical correlation analysis (Lai and Fyfe (1998), Tabanick and Fidell (1996)] can solve the above problem with additional norm and orthogonal constraints on translated features:

$$E\left\{\tilde{X}^T \cdot \tilde{X}\right\} = I \quad \text{and} \quad E\left\{\tilde{Y} \cdot \tilde{Y}\right\} = I \tag{14}$$

The CCA is described in further details in Hotelling (1936) and Hardoon et al(2004). The optimization criteria used for all three cross modal associations CFA, CCA and LSA

exhibit a high degree of noise tolerance. Hence the correlation features extracted perform better as compared to normal correlation analysis against noisy environmental conditions.

4. Hybrid audiovisual fusion

In this Section, we describe the fusion approach used for combing the extracted audio-lip correlated components with mutually independent audio and visual speech features.

4.1 Feature fusion of correlated components

The algorithm for fusion of audiovisual feature extracted using the cross modal association (CMA) models (a common term being used here to represent LSA, CFA or CCA analysis methods) can be described as follows:

Let f_A and f_L represent the audio MFCC and lip-region eigenlip features respectively. A and B represent the CMA transformation matrices (LSA, CFA or CMA matrices). One can apply CMA to find two new feature sets $f_A^{'} = A^T f_A$ and $f_L^{'} = B^T f_L$ such that the between-class cross-modal association coefficient matrix of $f_A^{'}$ and $f_L^{'}$ is diagonal with maximised diagonal terms. However, maximised diagonal terms do not necessarily mean that all the diagonal terms exhibit strong cross modal association. Hence, one can pick the maximally correlated components that are above a certain correlation threshold θ_k. Let us denote the projection vector that corresponds to the diagonal terms larger than the threshold θ_k by $\widetilde{w}_A$. Then the corresponding projections of f_A and f_L are given as:

$$\tilde{f}_A = \tilde{w}_A^T . f_A \text{ and } \tilde{f}_A = \tilde{w}_L^T . f_A \tag{15}$$

Here $\tilde{f}_A$ and $\tilde{f}_L$ are the correlated components that are embedded in f_A and f_L. By performing feature fusion of correlated audio and lip components, we obtained the CMA optimised feature fused audio-lip feature vector:

$$\tilde{f}_{AL}^{LSA} = \begin{bmatrix} \tilde{f}_A^{LSA} & \tilde{f}_L^{LSA} \end{bmatrix} \tag{16}$$

$$\tilde{f}_{AL}^{CFA} = \begin{bmatrix} \tilde{f}_A^{CFA} & \tilde{f}_L^{CFA} \end{bmatrix} \tag{17}$$

$$\tilde{f}_{AL}^{CCA} = \begin{bmatrix} \tilde{f}_A^{CCA} & \tilde{f}_L^{CCA} \end{bmatrix} \tag{18}$$

4.2 Late fusion of mutually independent components

In the Bayesian framework, late fusion can be performed using the product rule assuming statistically independent modalities, and various methods have been proposed in the literature as alternatives to the product rule such as max rule, min rule and the reliability-based weighted summation rule (Nefian et al(2002), Movellan and Mineiro(1997)). In fact, the most generic way of computing the joint scores can be expressed as a weighted summation

$$\rho(\lambda_r) = \sum_{n=1}^{N} w_n \log P(f_n \mid \lambda_r) \quad for\ r = 1, 2, \ldots, R \tag{19}$$

where $\rho_n(\lambda_r)$ is the logarithm of the class-conditional probability, $P(f_n|\lambda_r)$, for the n^{th} modality f_n given class λ_r, and w_n denotes the weighting coefficient for modality n, such that $\Sigma_n w_n = 1$. Then the fusion problem reduces to a problem of finding the optimal weight coefficients. Note that when $w_n = \frac{1}{N} \quad \forall n$, Eqn. 14 is equivalent to the product rule. Since the w_n values can be regarded as the reliability values of the classifiers, this combination method is also referred to as RWS (Reliability Weighted Summation) rule (Jain et al(2005), Nefian et al(2002)).The statistical and the numerical range of these likelihood scores vary from one classifier to another. Using sigmoid and variance normalization as described in (Jain et al(2005)), the likelihood scores can be normalised to be within the (0, 1) interval before the fusion process.

The hybrid audiovisual fusion vector in this Chapter was obtained by late fusion of feature fused correlated components ($\tilde{f}_{AL}^{LSA}$, $\tilde{f}_{AL}^{CFA}$, $\tilde{f}_{AL}^{CCA}$) with uncorrelated and mutually independent implicit lip texture features, and audio features with weights selected using the an automatic weight adaptation rule and is described in the next Section.

4.3 Automatic weight adaptation

For the RWS rule, the fusion weights are chosen empirically, whereas for the automatic weight adaptation, a mapping needs to be developed between modality reliability estimate and the modality weightings. The late fusion scores can be fused via sum rule or product rule. Both methods were evaluated for empirically chosen weights, and it was found that the results achieved for both were similar. However, sum rule for fusion has been shown to be more robust to classifier errors in literature (Jain et al (2005), Sanderson (2008)), and should perform better when the fusion weights are automatically, rather than empirically determined. Hence the results for additive fusion only, are presented here. Prior to late fusion, all scores were normalised to fall into the range of [0,1], using min-max normalisation.

$$\begin{aligned} P\left(S_i \left| x_A, x_V \right.\right) &= \alpha P(S_i \left| x_A \right.) + \beta P(S_i \left| x_v \right.) \\ P\left(S_i \left| x_A, x_V \right.\right) &= P(S_i \left| x_A \right.)^{\alpha} \times P(S_i \left| x_v \right.)^{\beta} \end{aligned} \tag{20}$$

where

$$\alpha = \begin{Bmatrix} 0 & c \le 1 \\ 1+c & -1 < c < 0 \\ 1 & c \ge 0 \end{Bmatrix} \quad and \quad \beta = \begin{Bmatrix} 1 & c \le 0 \\ 1-c & 0 < c < 1 \\ 0 & c \ge 1 \end{Bmatrix} \tag{21}$$

where, x_A and x_V refer to the audio test utterance and visual test sequence/image respectively.

To carry out automatic fusion, that adapts to varying acoustic SNR conditions, a single parameter *c*, the *fusion parameter*, was used to define the weightings; the audio weight α and the visual weight β, i.e., both α and β dependent on *c*. Higher values of *c* (>*0*) place more emphasis on the audio module whereas lower values (<*0*) place more emphasis on the visual module. For $c \geq 1$, $\alpha = 1$ and $\beta = 0$, hence the audiovisual fused decision is based entirely on the audio likelihood score, whereas, for $c \leq -1$, $\alpha = 0$ and $\beta = 1$, the decision is based entirely on the visual score. So in order to account for varying acoustic conditions, only *c* has to be adapted.

The reliability measure was the audio log-likelihood score $\rho_n(\lambda_r)$. As the audio SNR decreases, the absolute value of this reliability measure decreases, and becomes closer to the threshold for client likelihoods. Under clean test conditions, this reliability measure increases in absolute value because the client model yields a more distinct score. So, a mapping between ρ and *c* can automatically vary α and β and hence place more or less emphasis on the audio scores. To determine the mapping function *c*(ρ), the values of *c* which provided for optimum fusion, c_{opt}, were found by exhaustive search for the *N* tests at each SNR levels. This was done by varying *c* from *–1* to *+1*, in steps of *0.01*, in order to find out which *c* value yielded the best performance. The corresponding average reliability measures were calculated, ρ_{mean}, across the *N* test utterances at each SNR level.

$$c(\rho) = c_{os} + \frac{h}{\exp\left[d.(\rho + \rho_{os})\right]} \tag{22}$$

A sigmoid function was employed to provide a mapping between the c_{opt} and the ρ_{mean} values, where c_{os} and ρ_{os} represent the offsets of the fusion parameter and reliability estimate respectively; *h* captures the range of the fusion parameter; and *d* determines the steepness of the sigmoid curve. The sigmoidal parameters were determined empirically to give the best performance. Once the parameters have been determined, automatic fusion can be carried out. For each set of *N* test scores, the ρ value was calculated and mapped to *c*, using *c* = *c*(ρ), and hence, α and β can be determined. This fusion approach is similar to that used in (Sanderson(2008)) to perform speech recognition. The method can also be considered to be a secondary classifier, where the measured ρ value arising from the primary audio classifier is *classified* to a suitable *c* value; also, the secondary classifier is *trained* by determining the parameters of the sigmoid mapping.

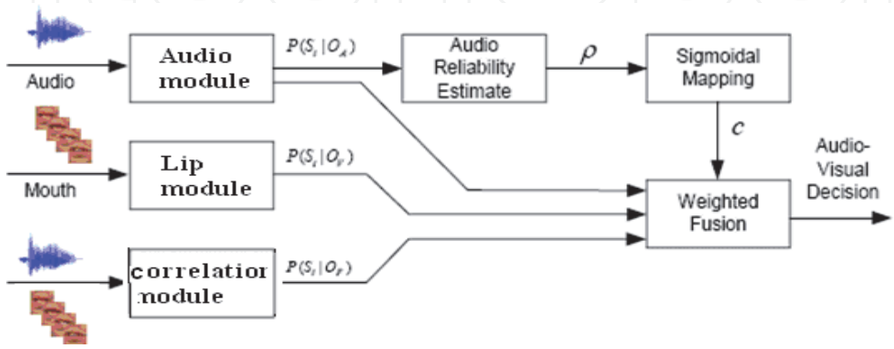

Fig. 1. System Overview of Hybrid Fusion Method

The described method can be employed to combine any two modules. It can also be adapted to include a third module. We assume here that only the audio signal is degraded when testing, and that the video signal is of fixed quality. The third module we use here is an audio-lip correlation module, which involves a cross modal transformation of feature fused audio-lip features based on CCA, CFA or LSA cross modal analysis as described in Section 3.
An overview of the fusion method described is given in Figure 1. It can be seen that the reliability measure, ρ, depends only on the audio module scores. Following the sigmoidal mapping of ρ, the fusion parameter c is passed into the fusion module along with the three scores arising from the three modules; fusion takes place to give the audiovisual decision.

5. Data corpora and experimental setup

A experimental evaluation of proposed correlation features based on cross-modal association models and their subsequent hybrid usion was carried out with two different audio-visual speaking face video corpora VidTIMIT (Sanderson(2008)) and (DaFEx (Battocchi et al (2004), Mana et al (2006)). Figure 2 show some images from the two corpora. The details of the two corpora are given in VidTIMIT (Sanderson(2008), DaFEx (Battocchi et al (2004), Mana et al (2006)).
The pattern recognition experiments with the data from the two corpora and the correlation features extracted from the data involved two phases, the training phase and the testing phase. In the training phase a 10-mixture Gaussian mixture model λ of a client's audiovisual feature vectors was built, 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 $log\ p(X|\lambda)$ of the time sequences X of audiovisual feature vectors under the usual assumption of statistical independence of successive feature vectors.
For testing replay attacks, we used a two level testing, a different approach from traditional impostor attacks testing used in identity verification experiments. Here the impostor attack is a surreptitious replay of previously recorded data and such an attack can be simulated by synthetic data. Two different types of replay attacks with increasing level of sophistication and complexity were simulated: the "*static*" replay attacks and the "*dynamic*" replay attacks.

(a) VidTIMIT corpus (b) DaFeX corpus

Fig. 2. Sample Images from VidTIMIT and DaFeX corpus

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 $log\ p(X'|\lambda)$ 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, an efficient photo-realistic audio-driven facial animation technique with near-perfect lip-synching of the audio and several image key-frames of the speaking face video sequence was done to create a artificial speaking character for each person (Chetty and Wagner(2008), Sanderson(2008).

In Bayesian framework, the liveness verification task can be essentially considered as a two class decision task, distinguishing the test data as a genuine client or an impostor. The impostor here is a fraudulent replay of client specific biometric data. For such 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 can be measured in terms of False Acceptance Rate (FAR) and False Reject Rate (FRR), as defined as (Eqn. 23):

$$FAR\ \% = \frac{I_A}{I_T} \times 100\ \% \qquad FRR\ \% = \frac{C_R}{C_T} \times 100\ \% \tag{23}$$

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 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. 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 a priori 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 subsets of data from the VidTIMIT and DaFeX were used. The gender-specific universal background models (UBMs) were developed using the training data from two sessions, Session 1 and Session 2, of the VidTIMIT corpus, and for testing Session 3 was used. Due to the type of data available (test session sentences differ from training session sentences), only text-independent speaker verification experiments could be performed with VidTIMIT. 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 the DaFeX database, similar gender-specific universal background models (UBMs) were obtained using training data from the text-dependent subsets corresponding to neutral expression. Ten sessions of the male and female speaking face data from these subsets were used for training and 5 sessions for testing.

For all the experiments, the global threshold was set using test data. For the male only subset of the 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 the male only subset for DaFeX 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 DaFeX subset, there were similar numbers of the client and impostor trials as in the male subset as we used 5 male and 5 female speakers from different subsets.

Different sets of experiments were conducted to evaluate the performance of the proposed correlation features based on cross modal association models (LSA, CCA and CMA), and their subsequent fusion in terms of DET curves and equal error rates (EER). Next Section discusses the results from different experiments.

6. Experimental results

Figure 3 plots the maximised diagonal terms of the between class correlation coefficient matrix after the LSA, CCA and CFA analysis of audio MFCC and lip-texture (f^{N}_{eigLip}) features. Results for the CFA analysis technique for the VidTIMIT male subset are only discussed here. As can be observed from Figure 3, the maximum correlation coefficient is around 0.7 and 15 correlation coefficients out of 40 are higher than 0.1.

Table 1 presents the EER performance of the feature fusion of correlated audio-lip fusion features (cross modal features) for varying correlation coefficient threshold θ. Note that, when all the 40 transformed coefficients are used, the EER performance is 6.8%. The EER performance is observed to have a minimum around 4.7% for threshold values from 0.1 to 0.4. The optimal threshold that minimises the EER performance and the feature dimension is found to be 0.4.

	EER(%) at (θ, dim)						
Θ	0.0	0.1	0.2	0.3	0.4	0.5	0.6
Dim	40	15	12	10	8	6	4
$\tilde{f}_{AL}^{CFA}$	6.8	4.7	5.3	5.0	4.7	7.4	10.3
$\tilde{f}_{AL}^{CCA}$	7.5	5.18	5.84	5.5	5.18	8.16	11.36
$\tilde{f}_{AL}^{LSA}$	11.7	8.09	9.12	8.6	8.09	12.74	17.74

Table 1. Results for correlation features based in CMA models: EERs at varying correlation coefficient threshold values (θ) with the corresponding projection dimension (dim)

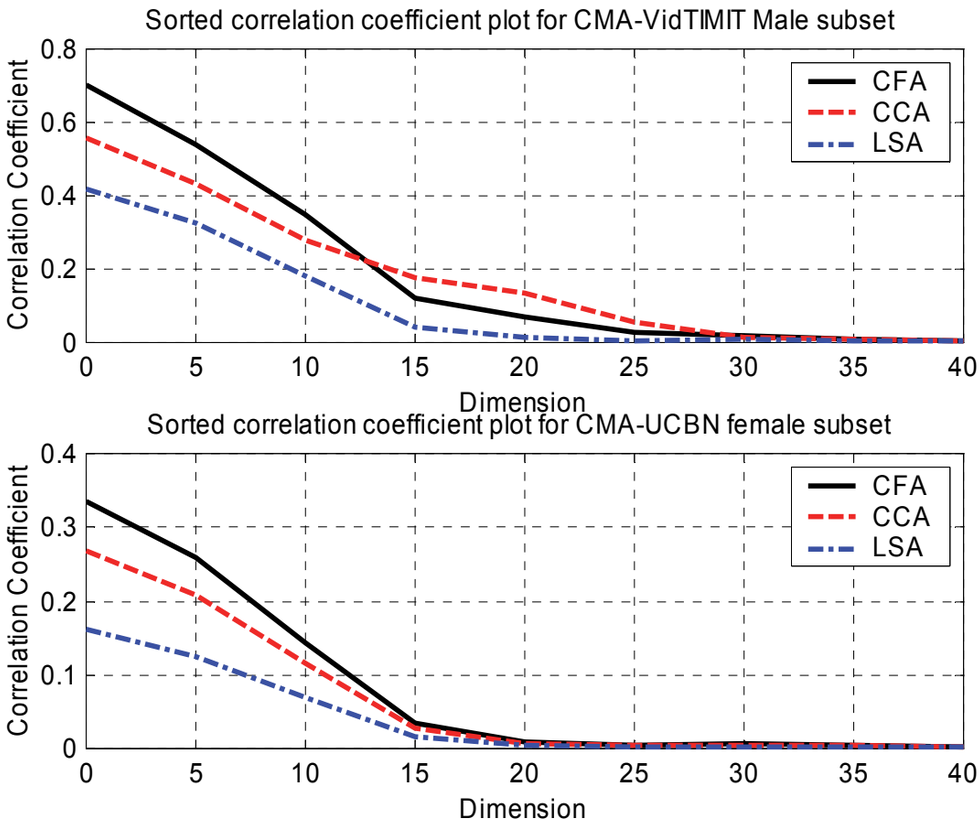

Fig. 3. Sorted correlation coefficient plot for audio and lip texture cross modal analysis

As can seen in Table 2 and Figure 4, for static replay attack scenarios (from the last four rows in Table 2), the nonlinear correlation components between acoustic and orafacial articulators during speech production is more efficiently captured by hybrid fusion scheme involving late fusion of audio f_{mfcc} features, f_{eigLip} lip features, and feature-level fusion of correlated audio-lip $\tilde{f}_{mfcc-eigLip}$ features). This could be due to modelling of identity specific mutually independent, loosely coupled and closed coupled audio-visual speech components with this approach, resulting in an enhancement in overall performance.

Modality	VidTIMIT male subset			DaFeX male subset		
	CFA EER (%)	CCA EER (%)	LSA EER (%)	CFA EER (%)	CCA EER (%)	LSA EER (%)
f_{mfcc}	4.88	4.88	4.88	5. 7	5. 7	5. 7
f_{eigLip}	6.2	6.2	6.2	7.64	7.64	7.64
$f_{mfcc-eigLip}$	7.87	7.87	7.87	9.63	9.63	9.63
$\tilde{f}_{mfcc-eigLip}$	3.78	2.3	2.76	4.15	2.89	3.14
$f_{mfcc} + f_{mfcc-eigLip}$	2.97	2.97	2.97	3.01	3.01	3.01
$f_{mfcc} + \tilde{f}_{mfcc-eigLip}$	0.56	**0.31**	0.42	0.58	0.38	0.57
$f_{mfcc} + f_{eigLip} + f_{mfcc-eigLip}$	6.68	6.68	6.68	7.75	7.75	7.75
$f_{mfcc} + f_{eigLip} + \tilde{f}_{mfcc-eigLip}$	0.92	0.72	0.81	0.85	0.78	0.83

Table 2. EER performance for static replay attack scenario with late fusion of correlated components with mutually independent components: (+) represents RWS rule for late fusion, (-) represents feature level fusion)

Though all correlation features performed well, the CCA features appear to be the best performer for static attack scenario, with an EER of 0.31%. This was the case for all the subsets of data shown in Table 2. Also, the EERs for hybrid fusion experiments with $\tilde{f}_{mfcc-eigLip}$ correlated audio lip features performed better as compared to ordinary feature fusion of $f_{mfcc-eigLip}$ features. EERs of 0.31% and 0.72% were achieved for $f_{mfcc} + \tilde{f}_{mfcc-eigLip}$ and $f_{mfcc} + f_{eigLip} + \tilde{f}_{mfcc-eigLip}$ features, the hybrid fusion types involving CMA optimised correlated features, as compared to an EER of 2.97% for $f_{mfcc} + f_{mfcc-eigLip}$ features and 6.68% for $f_{mfcc} + f_{eigLip} + f_{mfcc-eigLip}$ features, which are hybrid fusion types based on ordinary feature fusion of uncorrelated audio-lip features. This shows that correlation features based on proposed cross-modal association models can extract the intrinsic nonlinear temporal correlations between audio-lip features and could be more useful for checking liveness.
The EER table in Table 3 shows the evaluation of hybrid fusion of correlated audio-lip features based on cross modal analysis (CFA, CCA and LSA) for dynamic replay attack scenario. As can be seen, the CMA optimized correlation features perform better as compared to uncorrelated audio-lip features for complex dynamic attacks. Further, for the VidTIMIT male subset, it was possible to achieve the best EER of l0.06% for $f_{mfcc} + f_{eigLip} + \tilde{f}_{mfcc-eigLip}$ features, a hybrid fusion type involving feature fusion of correlated audio-lip features based on CCA analysis.

7. Conclusion

In this Chapter, we have proposed liveness verification for enhancing the robustness of biometric person authentication systems against impostor attacks involving fraudulent replay of client data. Several correlation features based on novel cross-modal association models have been proposed as an effective countermeasure against such attacks. These new

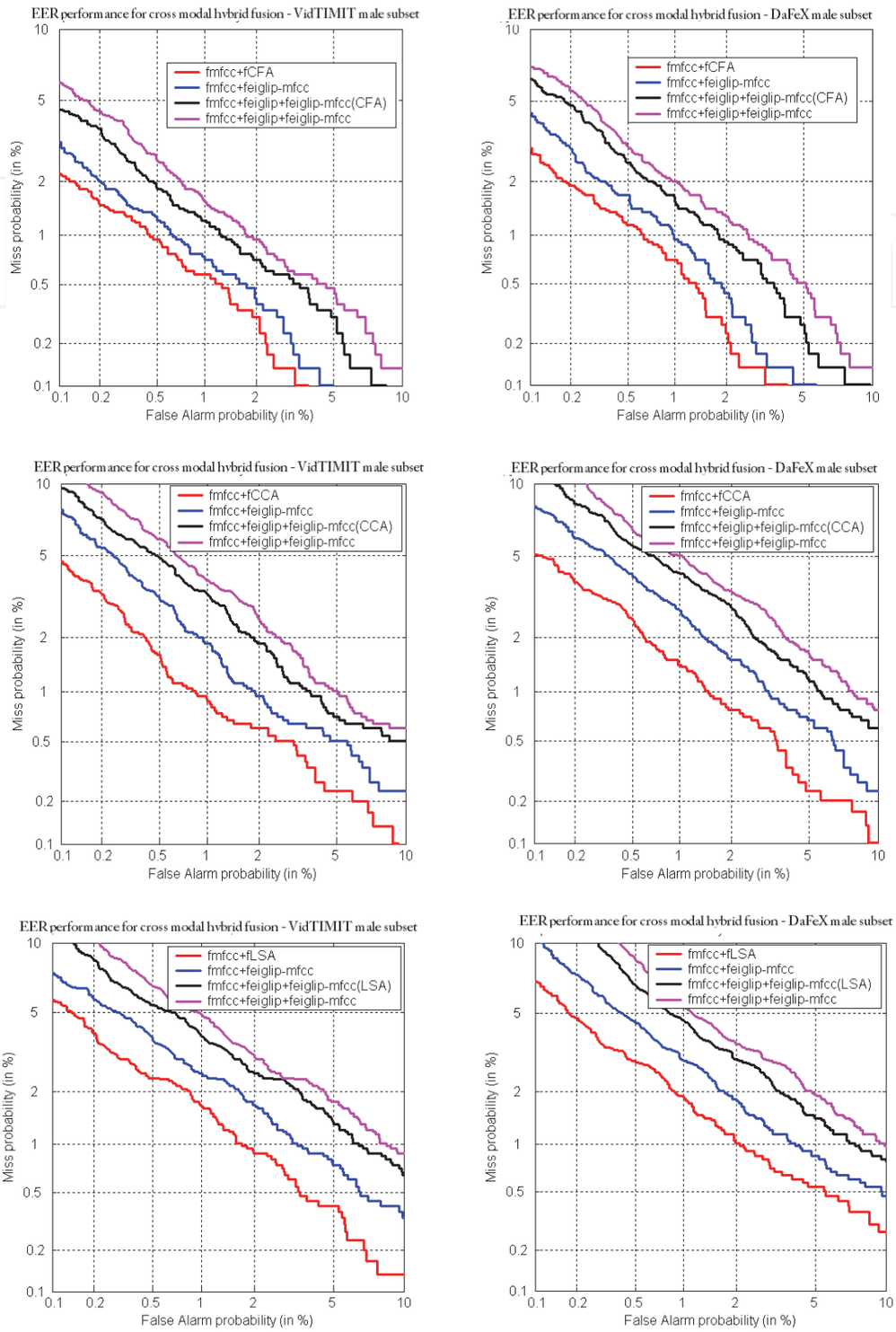

Fig. 4. DET curves for hybrid fusion of correlated audio-lip features and mutually independent audio-lip features for static replay attack scenario

correlation measures model the nonlinear acoustic-labial temporal correlations for the speaking faces during speech production, and can enhance the system robustness against replay attacks.

Further, a systematic evaluation methodology was developed, involving increasing level of difficulty in attacking the system – moderate and simple static replay attacks, and, sophisticated and complex dynamic replay attacks, allowing a better assessment of system vulnerability against attacks of increasing complexity and sophistication. For both static and dynamic replay attacks, the EER results were very promising for the proposed correlation features, and their hybrid fusion with loosely coupled (feature-fusion) and mutually independent (late fusion) components, as compared to fusion of uncorrelated features. This suggests that it is possible to perform liveness verification in authentication paradigm. and thwart replay attacks on the system. Further, this study shows that, it is difficult to beat the system, if underlying modelling approach involves efficient feature extraction and feature selection techniques, that can capture intrinsic biomechanical properties accurately.

	VidTIMIT male subset			DaFeX male subset		
Modality	CFA EER (%)	CCA EER (%)	LSA EER (%)	CFA EER (%)	CCA EER (%)	LSA EER (%)
f_{eigLip}	36.58	36. 58	36. 58	37.51	37. 51	37. 51
$f_{mfcc-eigLip}$	27.68	27.68	27.68	28.88	28.88	28.88
$\tilde{f}_{mfcc-eigLip}$	24.48	22.36	23.78	26.43	24.67	25.89
$f_{mfcc} + f_{mfcc-eigLip}$	22.45	22.45	22.45	23.67	23.67	23.67
$f_{mfcc} + \tilde{f}_{mfcc-eigLip}$	17.89	16.44	19.48	18.46	17.43	20.11
$f_{mfcc} + f_{eigLip} + f_{mfcc-eigLip}$	21.67	21.67	21.67	25.42	25.42	25.42
$f_{mfcc} + f_{eigLip} + \tilde{f}_{mfcc-eigLip}$	14.23	10.06	12.27	16.68	12.36	13.88

Table 3. EER performance for dynamic replay attack scenario with late fusion of correlated components with mutually independent components

However, though the EER performance appeared to be very promising for static replay attack scenarios (EER of 0.31 % for CCA features), the deterioration in performance for more sophisticated - dynamic replay attack scenario (EER of 10.06 % for CCA features), suggests that, there is an urgent need to investigate more robust feature extraction, feature selection, and classifier approaches, as well as sophisticated replay attack modelling techniques. Further research will focus on these two aspects.

8. References

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

[2] Chaudhari U.V, Ramaswamy G.N, Potamianos G, and Neti C.(2003) 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.

[3] Chibelushi C.C, Deravi F, and Mason J(2002) A Review of Speech-Based Bimodal Recognition. IEEE Transactions on Multimedia, 4(1):23–37.

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

[5] Fisher III J. W, Darrell T, Freeman W. T, Viola P (2000), Learning joint statistical models for audio-visual fusion and segregation, Advances in Neural Information Processing Systems (NIPS), pp. 772-778.

[6] 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.

[7] Goecke R. and Millar J.B.(2003). 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.

[8] Gurbuz S, Tufekci Z, Patterson T, and Gowdy J.N (2002) 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.

[9] Hershey J. and Movellan J (1999) Using audio-visual synchrony to locate sounds, Proc. Advances in Neural Information Processing Systems (NIPS), pp. 813-819.

[10] Hotelling H (1936) Relations between two sets of variates Biometrika, 28:321 377.

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

[12] Jain A, Nandakumar K, and Ross A (2005) Score Normalization in Multimodal Biometric Systems, Pattern Recognition.

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

[14] Lai P. L., and Fyfe C (1998), Canonical correlation analysis using artificial neural networks, Proc. European Symposium on Artificial Neural Networks (ESANN).

[15] Li M, Li D, Dimitrova N, and Sethi I.K(2003), Audio-visual talking face detection, Proc. International Conference on Multimedia and Expo (ICME), pp. 473-476, Baltimore, MD.

[16] Li D, Wei G, Sethi I. K, Dimitrova N(2001), Person Identification in TV programs, Journal on Electronic Imaging, Vol. 10, Issue. 4, pp. 930-938.

[17] Liu X, Liang L, Zhaa Y, Pi X, and Nefian A.V(2002) Audio-Visual Continuous Speech Recognition using a Coupled Hidden Markov Model. In Proc. International Conference on Spoken Language Processing.

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

[19] Mana N, Cosi P, Tisato G, Cavicchio F, Magno E. and Pianesi F(2006) 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.

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

[21] Movellan, J. and Mineiro, P(1997), "Bayesian robustification for audio visual fusion". In Proceedings of the Conference on Advances in Neural information Processing Systems 10 (Denver, Colorado, United States). M. I. Jordan, M. J. Kearns, and S. A. Solla, Eds. MIT Press, Cambridge, MA, 742-748.

[22] Nefian V, Liang L. H, Pi X, Liu X, and Murphy K. (2002) Dynamic Bayesian Networks for Audio-visual Speech Recognition, EURASIP Journal on Applied Signal Processing, pp. 1274-1288.

[23] Pan H, Liang Z, and Huang T(2000)A New Approach to Integrate Audio and Visual Features of Speech. In Proc. IEEE International Conference on Multimedia and Expo., pages 1093 – 1096.

[24] Potamianos G, Neti C, Luettin J, and Matthews I (2004) Audio-Visual Automatic Speech Recognition: An Overview. Issues in Visual and Audio-Visual Speech Processing.

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

[26] Tabachnick B, and Fidell L. S (1996), Using multivariate statistics, Allyn and Bacon Press.

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

Advanced Biometric Technologies

Edited by Dr. Girija Chetty

ISBN 978-953-307-487-0
Hard cover, 382 pages
Publisher InTech
Published online 09, August, 2011
Published in print edition August, 2011

The methods for human identity authentication based on biometrics â€“ the physiological and behavioural characteristics of a person have been evolving continuously and seen significant improvement in performance and robustness over the last few years. However, most of the systems reported perform well in controlled operating scenarios, and their performance deteriorates significantly under real world operating conditions, and far from satisfactory in terms of robustness and accuracy, vulnerability to fraud and forgery, and use of acceptable and appropriate authentication protocols. To address some challenges, and the requirements of new and emerging applications, and for seamless diffusion of biometrics in society, there is a need for development of novel paradigms and protocols, and improved algorithms and authentication techniques. This book volume on â€œAdvanced Biometric Technologiesâ€ is dedicated to the work being pursued by researchers around the world in this area, and includes some of the recent findings and their applications to address the challenges and emerging requirements for biometric based identity authentication systems. The book consists of 18 Chapters and is divided into four sections namely novel approaches, advanced algorithms, emerging applications and the multimodal fusion. The book was reviewed by editors Dr. Girija Chetty and Dr. Jucheng Yang We deeply appreciate the efforts of our guest editors: Dr. Norman Poh, Dr. Loris Nanni, Dr. Jianjiang Feng, Dr. Dongsun Park and Dr. Sook Yoon, as well as a number of anonymous reviewers.

How to reference

In order to correctly reference this scholarly work, feel free to copy and paste the following:

Girija Chetty and Emdad Hossain (2011). Multimodal Fusion for Robust Identity Authentication: Role of Liveness Checks, Advanced Biometric Technologies, Dr. Girija Chetty (Ed.), ISBN: 978-953-307-487-0, InTech, Available from: http://www.intechopen.com/books/advanced-biometric-technologies/multimodal-fusion-for-robust-identity-authentication-role-of-liveness-checks

INTECH
open science | open minds

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821

IntechOpen

IntechOpen