Robust face-voice based speaker identity verification using multilevel fusion

Girija Chetty *, Michael Wagner

Human Computer Communication Laboratory, School of Information Sciences and Engineering, University of Canberra, LPO Box 340, Building 11, UC, Bruce, ACT 2601, Australia

**Abstract**

In this paper, we propose a robust multilevel fusion strategy involving cascaded multimodal fusion of audio–lip–face motion, correlation and depth features for biometric person authentication. The proposed approach combines the information from different audio–video based modules, namely: audio–lip motion module, audio–lip correlation module, 2D + 3D motion–depth fusion module, and performs a hybrid cascaded fusion in an automatic, unsupervised and adaptive manner, by adapting to the local performance of each module. This is done by taking the output-score based reliability estimates (confidence measures) of each of the module into account. The module weightings are determined automatically such that the reliability measure of the combined scores is maximised. To test the robustness of the proposed approach, the audio and visual speech (mouth) modalities are degraded to emulate various levels of train/test mismatch; employing additive white Gaussian noise for the audio and JPEG compression for the video signals. The results show improved fusion performance for a range of tested levels of audio and video degradation, compared to the individual module performances. Experiments on a 3D stereovision database AVOZES show that, at severe levels of audio and video mismatch, the audio, mouth, 3D face, and tri-module (audio–lip motion, correlation and depth) fusion EERs were 42.9%, 32%, 15%, and 7.3%, respectively, for biometric person authentication task.

**Keywords:**

Lip
3D Face
Voice
Biometric
Identity verification
Robust
Multilevel fusion

1. Introduction

Biometrics is a field of security technology devoted to verification or identification of individuals using physiological or behavioural traits. Verification, a binary classification problem, involves the validation of a claimed identity whereas identification, a multi-class problem, involves identifying a user from a set of enrolled subjects; and becomes more difficult as the number of enrollees increases.

Most of the speaker recognition systems currently deployed are based on modelling a speaker based on unimodal information, i.e., either audio or visual features. Audio-based identification achieves high performance when the signal-to-noise ratio (SNR) is high. Yet, the performance degrades quickly as the test SNR decreases (referred to as a train/test mismatch), as shown in [1] and elsewhere. Using visual modality in addition to voice information, such as 3D face or 2D region around mouth can make the system robust against SNR degradation, typical of mismatch between and training and test operating environment. However, visual modality based speaker modelling approaches are also susceptible to pose/illumination variation, occlusion, and poor image quality [1,3]. Further, use of 2D visual speech features extracted from mouth region on its own cannot model a speaking face in its entirety, and normally used along with other biometric modalities. However, mouth region contains important liveness related information, which can be used to detect fraudulent replay attacks involving a still photo of the speaker and replay of audio, or artificially synthesized speaking face.

To combat these limitations of unimodal modules, a multimodal fusion approach based on speaking face video sequences is proposed here. The proposed approach allows significant improvement in performance and robustness against impostor attacks and fraudulent replay attacks for biometric person authentication task.

The audio, face, and mouth modalities contain non-redundant, complementary information about identity of the person. In addition, audio and lip motion during speech production are partially correlated, comprising mutually dependent and mutually independent components. Further, use of 3D face dynamics in addition, allows better speaker models to be built [4,5], as we can better quantify the differences between two persons’ facial feature variations in 3D as compared to 2D face images. The subtle nuances related to facial expressions and gestures during speaking act that can best discriminate individuals can also be modelled better with 3D face dynamics. From a biometrics point of view, the concept of recognizing a person based on 3D facial motion during speech is attractive; since facial movements comprise a complex sequence of muscle activations, and it is almost impossible to imitate another person’s facial speech and expressions, as these characteristics are unique to an individual [6,7]. In experimental psychology, determining the precise role of 3D facial motion in ascertaining identity is still largely unknown, and is being actively pursued.
This paper is organised as follows. The next section describes the motivation for using proposed multilevel multimodal fusion. Section 3 and 4 describe the proposed fusion approach for building speaking face models. In Section 5, the stereovision audio visual corpus AVOZES used for evaluation is described. In Section 6, the experimental results of extensive evaluations examining the individual module performance and multilevel fusion performance for a biometric security application for a specific scenario namely Speaker Identity Verification (SIV) scenario. The results are discussed in Section 7, and finally in references, conclusions from the results are drawn.

2. Multilevel speaking face modelling

This section discusses the motivation for using proposed multimodal fusion approach for robust speaking face modelling based on some recent findings in cognitive psychology [6] and psycho-physical analysis of visual speech [7]. As with the other forms of biological motion, humans are known to be very sensitive to the realism in the ways the lips move. One of the most significant findings by Yehia, Kuratate, Munhall, and Bateson [8,9] suggest that in order to determine the elements that come to play during analysis of visual speech, it is important to capture the detailed 3D deformations of faces when talking [9]. Yehia, Bateson, and Kuratate [8] suggest that a speaking face is a kinematic-acoustic system in motion, and the shape, texture and acoustic features during speech production are correlated in a complex way, and a single neuromotor source controlling the vocal tract behaviour is responsible for both the acoustic and the visible attributes of speech production. Hence, for a speaking face not only the facial motion and speech acoustics, there is an indirect anatomical coupling created by the complex of strap muscles running between the floor of the mouth, through the thyroid bone, attaching to the outer edge of the cricothyroid cartilage, as shown in Fig. 1. Due to this indirect coupling, speakers tend to raise the pitch when head goes up while talking.

It has also been shown by several other linguistic and psycho-physical researchers [6–9], that the facial movements play an important role in interpreting spoken conversations and emotions. They occur continuously during social interactions and conversations. They include lip movements when talking, conversational signals, emotion displays and manipulators to satisfy biological needs. Unfortunately, when and how a movement appears and disappears, and how co-occurrent movements are integrated (co-articulation effects, for instance) are difficult to quantify.

In addition, the problem of overlaying and blending facial movements in time, and the way felt emotions are expressed in facial activity during speech, have not received much attention. This suggests that during speech production other regions of the face in addition to the lip region are active, and the activities of human facial muscles for this act are far from simply additive.

A typical example would be smiling while speaking. The Zygomatic Major and Minor muscles contract to pull the corner of lip outward, resulting in a smile. The viseme corresponding to the diphthong/oU/ in the word “Hello” requires the contraction of the lip funneler Orbicularis Oris, which drives the lips into a tight, pursed shape. However, the activation of the Zygomatic Major and Minor muscles together with the lip funneler Orbicularis Oris would create an unnatural movement. The activation of a muscle may require the deactivation of other muscles in the jaw and chin region.

These findings from face speech anatomy provide clues that facial movements during speech involve highly complex biomechanics with depth, motion and correlation interactions. Capturing these interactions can truly enhance the performance of face modelling approaches for complex application scenarios such as biometric security systems. These complex spatio-temporal correlations it appears can be modelled better with multilevel fusion strategy involving cascaded fusion of multimodal (2D + 3D-acoustic, lip, face) modules. Such a multilevel fusion strategy as shown in investigations reported here can truly establish the identity of the speaker, distinguish it from an impostor or a fraudulent replay, and be robust to degradation in audio and visual operating conditions.

The proposed multilevel spatio-temporal face modelling approach thus captures the correlated and mutually independent components of facial movements during speech between acoustic-labial articulators as well as other areas of face and head such as jaw, chin, forehead, and eyebrows.

3. Fusion strategy

The proposed fusion strategy based on multilevel fusion of acoustic, 2D and 2D face modules is described in this section.

3.1. Acoustic features module

The MFCC features (mel frequency cepstral coefficients) of dimension 16 were extracted from each frame. Mel frequency cepstral coefficients (MFCCs) are coefficients that represent audio
[3,18], and more commonly used for speaker recognition tasks. They are derived from a type of cepstral representation of the audio clip (“spectrum-of-a-spectrum”). The difference between the cepstrum and the mel frequency cepstrum is that in the MFC, the frequency bands are positioned logarithmically (on the mel scale) which approximates the human auditory system’s response more closely than the linearly spaced frequency bands obtained directly from the FFT or DCT. This can allow for better modelling of vocal tract transfer function. MFCCs are commonly derived as directly from the FFT or DCT. This can allow for better modelling of vocal tract transfer function. MFCCs are commonly derived as

\[ f_{GRD} = \left[ f'_1, f'_2, f'_3, \ldots, f'_M \right] \]

where \( f'_k \) refers to the observation feature vector at grid points, respectively. The first M DCT coefficients along the zigzag scan order, both for x and y directions, are combined to form a feature vector \( f \) of dimension 2M as depicted in Fig. 2. This feature vector representing the dense grid motion will be denoted by \( f_{GRD} \) in rest of the paper.

For lip contour extraction, we employ the lip geometric key points and fit polynomials on the outer lip contour based on a technique proposed in [12,13]. The technique is based on six designated key points detected on the lip contour. The algorithm fits additional points on the outer lip key points by guiding a “jumping snake” onto the upper lip boundary [12]. The additional detected key points serve as the junction points of four cubic polynomials and two line segments to be fitted onto the lip contour via least squares optimization. The DCT coefficients computed separately for the x and y directions are concatenated to form the feature vector that is denoted by \( f_{CTR} \).

The three types of visual features were concatenated to form a 24 dimensional feature vector. This is shown in Eq. (3) where \( f_{DCT}, f_{GRD}, f_{CTR} \) represent the DCT, grid, and contour based lip motion features, respectively, and \( o_t \) refers to the observation feature vector for the frame at time \( t \).

\[ o_t = \left[ o'_t^{DCT}, o'_t^{GRD}, o'_t^{CTR} \right] \]

Similarly to the audio case, we have \( T_v \) visual observations (\( T_v \approx 2T_v \)) and a sequence, \( O_m \), of visual mouth speech feature vectors or observations denoted by:

\[ o_m = \{ o_1, o_2, \ldots, o_t, \ldots, o_T \} \]

3.3. Acoustic-2D face cross modal features module

For this module, we extract explicit correlation features based on cross modal factor (CMF) analysis technique [19], which allows extraction of the correlated components in audio and lip modalities. CMF features model the semantic correlation between visual lip region and the associated speech. The method consists of three major steps: the construction of a joint multimodal feature space, the normalization, and the singular value decomposition (SVD).

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

\[ X = [V_1, V_2, \ldots, V_m, A_1, \ldots, A_n] \]

where

\[ V_t = [v_t(1), v_t(2), \ldots, v_t(t)]^T \]

and

\[ A_t = [a_t(1), a_t(2), \ldots, a_t(t)]^T \]
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:

$$\hat{X}_j = \frac{X_j}{\max(\text{abs}(X_j))}$$

After normalization all elements in normalized matrix $\hat{X}$ have values between $-1$ and $1$. SVD can then be performed as follows:

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

where $S$ and $D$ are matrices composing of left and right singular vectors and $V$ is diagonal matrix of singular values in descending order. Keeping only the first and most important $k$ singular vectors in $S$ and $D$, we can derive an optimal approximation of $\hat{X}$ with reduced feature dimensions, where semantic (correlation) information between visual and audio features is mostly preserved.

The CMF features used in the work reported here involves casting the optimization problem above differently, allowing the extraction of optimal transformations that can best represent (or identify) the coupled patterns between the features of the two different subsets, under the linear correlation model. The optimization criterion to obtain the optimal transformations and hence the CMF features, is as described here:

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

$$||X - YB||^2_F,$$

where

$$A^T A = I \quad \text{and} \quad B^T B = I$$

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

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

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:

$$||X - YB||^2_F = \text{trace}((X - YB)^T(X - YB))$$

$$= \text{trace}(XAX^T + YBY^T - XAB^T Y^T - YBA^T X^T)$$

$$= \text{trace}(X^T X) + \text{trace}(Y^T Y) - 2 \cdot \text{trace}(XAB^T Y^T)$$

where 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 maximize $\text{trace}(XAB^T Y^T)$ will minimize Eq. (3). It can be shown that such matrices are given by:

$$\begin{cases} A = S_{xy} \\ B = D_{xy} \end{cases}$$

where

$$X^T Y = S_{xy} \cdot V_{xy} \cdot D_{xy}$$

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

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

Corresponding vectors in $\tilde{X}$ and $\tilde{Y}$ are thus optimized to represent the coupled relationships between the two feature subsets without being affected by distribution patterns within each subset.

For our problem, we apply CFA to audio and lip vectors to find two new feature sets $f_A^i = A_i f_A^j$ and $f_L^i = B_i f_L^j$ such that the between-class cross modal association coefficient matrix of $f_A^i$ and $f_L^j$ 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 $\theta$. Let us denote the projection vector that corresponds to the diagonal terms larger than the threshold $\theta$ by $\tilde{w}_A$ and $\tilde{w}_L$. Then the corresponding projections of $f_A$ and $f_L$ are given as:

$$\tilde{f}_A = \tilde{w}_A^T \cdot f_A$$

$$\tilde{f}_L = \tilde{w}_L^T \cdot f_L$$

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 obtain the CFA optimized feature fused audio–lip feature vector:

$$\tilde{f}_{AL} = [\tilde{f}_A \ \tilde{f}_L]$$

3.4. 3D Face features module

The 3D facial feature module is described in detail in [14, 16]. The dimensionality of the 3D features varies depending on the type of the facial data representation and feature extraction techniques. For 3D face module, each face is modelled with 3D shape and texture features, the TEX-GABOR for texture features [14, 15], and CURV-PD for the shape features [15, 16].

4. Multimodal fusion

4.1. Automatic adaptive fusion weight computation

The following design criteria were taken into account, when designing the proposed fusion strategy. Information from the acoustic, face, cross-modal and 3D face features should be combined in a multilevel fusion mode to capture the visual manifestation of speech in terms of closely coupled and mutually independent 2D and 3D audio-visual components. The system must be robust to mild through adverse test levels of both audio and visual speech (mouth) noise. The contribution from each source of information to the final decision must be weighted dynamically by taking the current reliability of each source of information into account. The module score weightings must be determined in an automatic unsupervised manner. Given these criteria, we decided to use cascaded fusion of audio–lip–face features in a hybrid fusion scheme. The hybrid fusion modules use a combination of late fusion and feature fusion of audio–lip motion, correlation and depth features, based on the theoretical and empirical evidence from findings in the previous related literature [20–22]. With regards to the type of fusion rule, the sum rule is known to be superior to the product rule [20, 22], particularly when the module scores have large errors. Thus, the weighted sum rule should be resilient to noise and is a good choice for score level fusion, particularly in this application where either or both of the audio and video (mouth) modalities may be highly degraded. In terms of score normalization, the min–max norm, while being straightforward to implement has been found to have comparable performance to more complicated normalization techniques.

By taking all of this into consideration, the proposed multilevel fusion system is based on weighted sum score fusion with min–max normalization. The multilevel fusion is implemented as follows. We first perform a fusion of two modules (e.g., audio with
lip motion features and audio-visual cross-modal features). Then this bi-module fusion is extended to include an additional third module—fusion of 3D face features with audio–lip motion features, thus yielding tri-module cascaded fusion at multiple levels involving audio–lip motion, audio–lip correlation, and 3D depth features. We use \( IL(m) \) to denote the confidence score output from the \( m \)th module representing the log-likelihood that the observation \( O_m \) was caused by the client model/template where \( m \) belongs to the set \{ALM, ALR, 3LM\}, with ALM, ALR, and 3LM representing the audio–lip motion (\( O_{ALM} \)), audio–lip correlation (\( O_{ALR} \)), and 2D lip motion–3D depth features extracted from entire face (\( O_{3LM} \)) module observations, respectively. Two considerations must be taken into account, when choosing the module weightings for weighted fusion:

1. Individual modules operating under matched testing conditions will have different performances, e.g., visual speech based identity verification generally underperforms audio based identity verification.
2. Under mismatched testing conditions, each module may be subjected to different levels of noise (ideally statistically independent). Thus, it is necessary to capture the relative local reliability of each module.

Weighting schemes based on consideration (1) with fixed weights, are termed non-adaptive fusion, and schemes based on considerations (1) and (2), which adapt dynamically to the performance of each module, are termed adaptive fusion. Non-adaptive weights are determined a priori based on the evaluation performance of each individual module and are fixed for all ensuing tests. The weights do not take the testing conditions into account and hence cannot adapt, when the confidence (reliability) of one modality is low. This strategy will work reasonably well if the train–test mismatch is low. However, it may result in catastrophic fusion if the train/test mismatch is high.

The fusion module should take the local testing conditions into account and adapt the fusion parameters appropriately.

Thus, we have the weighted sum rule for \( O_1, \ldots, O_N \):

\[
S(O_1, \ldots, O_N; \lambda) = \sum_{m=1}^{N} w_m S(O_m; \lambda) = w_1 S(O_1; \lambda) + w_2 S(O_2; \lambda)
\]

where, \( S(O_1, \ldots, O_N; \lambda) \) represents the combined likelihood that the observations were produced by the client \( \lambda \) and \( w_m \) is the weight of the \( m \)th module, subject to the constraints that \( \sum w_m = 1 \) and \( 0 \leq w_m \leq 1 \) for \( m = 1 \ldots N \). The modules can be taken, in any combination, from the set \{ALM, ALR, 3LM\}. Given that the weights \( w_m \) are variable, some sort of reliability measure must be devised, which takes the confidence associated with each module into account, and is used to determine the \( w_m \) values.

4.2. Reliability measures

Module reliability parameters can be calculated at the signal level or at the module score level. Signal based reliability measures are derived directly from the signal observations prior to feature extraction. Examples include estimations of the signal-to-noise ratio [23] and the degree of voicing (harmonicity index) [24]. These have the disadvantage of having no corresponding visual reliability measure. Audio only reliability measures have been employed in many audiovisual speech processing studies [24–26]. This is undesirable, as the integrity of the video signal is vulnerable, due to the high video bandwidth requirement and sensitivity to illumination conditions. Even if an observation signal is of high quality, the module may still give a misclassification if the model for the speaker is a poor representation.

On the other hand, the distribution of the set of module confidence scores contains information not only about the reliability of that module’s decision, but also about the integrity of the observation signal. The reliability measure should not be module specific, i.e., it should be equally applicable to audio, face, and mouth based modules. Taking these points into account, it is better to calculate the reliability measure based on the module scores, as this can quantify both a train/test mismatch and the confidence in the modules’ classification decision.

If the client scores of a module are much higher than the thresholds, then the confidence level is high. Otherwise, the confidence is low. Various other metrics exist, which can be used to capture this confidence information better. Examples include score entropy [24], score dispersion [24], score variance [29], cross classifier coherence [30], and score difference [31]. However, we just used the scores itself as the confidence measures. For a test observation vector \( O_m \), we have the set of normalised client scores \( \{S(O_{m1}), \ldots, S(O_{mn})\} \). The score difference, \( \xi \), between the client confidence score and impostor confidence score can then be calculated as:

\[
\xi = S(O_{m1}) - S(O_{m2})
\]

where \( \lambda \) and \( \lambda \) are the client and the impostor classes, respectively, and \( m \) denotes the module.

4.3. Reliability mapping

A mapping between the reliability estimates and the module weightings is required. In the previous literature, a sigmoidal mapping was used to map the reliability estimates to the fusion weights [1,24]. The sigmoid curve is monotonic and is bounded by the range zero to one. The parameters of the sigmoid curve require training, which is difficult when the amount of audiovisual data is small. A linear relationship between audio module weighting \( a \) and the SNR audio reliability estimate was assumed in [27] and [26]. A piece-wise linear mapping was used to map the test SNR values to \( a \) in [26]. This is unsuitable here, as we need to consider the video reliability also.

Considering the small amount of audiovisual training data generally available, just three training sentences of around 4 seconds duration in the current study, it was decided to use a non-learned approach to map the reliability estimates to the \( a \) values. This was carried out as follows:

1. For each specific SIV trial (user transaction), the system is presented with two module observations, \( O_1 \) and \( O_2 \).
2. The two modules each generate a set of normalised scores \( \{S(O_{11}), \ldots, S(O_{1n})\} \) and \( \{S(O_{21}), \ldots, S(O_{2n})\} \), and the reliability estimates \( \xi_1 \) and \( \xi_2 \) are calculated, as in Eq. (19).
3. The fusion parameter \( a_1 \) is varied from 0 to 1 in steps of 0.05. For each of these \( a_1 \) values, the module scores \( \{S(O_{11}), \ldots, S(O_{1n})\} \) and \( \{S(O_{21}), \ldots, S(O_{2n})\} \) are combined using Eq. (18) with \( a_1 = 1 - a_2 \), to give the combined set of scores \( \{S_{x1}, \ldots, S_{xn}\} \).
4. The combined score set is subsequently normalised as before, to give \( \{S(O_{11}), \ldots, S(O_{1n})\} \), and the combined score reliability estimate, denoted by \( \xi_{12} \), is calculated as in Eq. (19). \( \xi_{12} \) can be thought of as a linear weighted combination of the individual module reliabilities \( \xi_1 \) and \( \xi_2 \) because...
\[ \xi_{12} = S(O_1, O_2 | \lambda) - S(O_1, O_2 | \tilde{\lambda}) \]
\[ = a_1 S(O_1 | \lambda) + a_2 S(O_2 | \lambda) - a_1 S(O_1 | \tilde{\lambda}) - a_2 S(O_2 | \tilde{\lambda}) \]
\[ = a_1 \xi_1 + a_2 \xi_2 \]  
(20)

where \( \lambda \) and \( \tilde{\lambda} \) are the client and impostor classes, as before.

We choose the \( a_2 \) value that maximises \( \xi_{12} \) for the given test according to Eq. (21) to give the fusion parameters \( a_{2opt} \) and \( a_{1opt} = 1 - a_{2opt} \). The maximum \( \xi_{12} \) value should correspond to the combined scores of highest confidence, i.e., maximises the score separation between the client class and impostor class. Finally, we combine \( \{ S(O_1 | \lambda) \} \) and \( \{ S(O_2 | \lambda) \} \) as in Eq. (18) (using \( a_{1opt} \) and \( a_{2opt} \)), to form the combined score list \( \{ S_{12opt} \} \) which is used to make the final verification decision.

\[ a_{2opt} = \arg \max_{a_2 \in [0, 1]} \{ \xi_{12} | a_2 \} \]  
(21)

It should be noted that the above procedure allows the fusion weights to be determined automatically in an unsupervised manner. No assumptions are made here as to which modules are employed, i.e., \( O_1 / O_2 \) above can represent any of the face, mouth, or audio experts, or indeed any module in general.

To illustrate this procedure, Fig. 3 gives four examples of the specific case of fusing the scores arising from audio–lip motion and audio–lip correlation test observations (in general, any two modules could be employed here), i.e., \( O_1 = O_{ALM} \) and \( O_2 = O_{ALR} \). For Fig. 3a, the audio–lip motion and audio–lip correlation score reliability estimate values, \( \xi_1 \) and \( \xi_2 \) are 0.74 and 0.11, respectively. The variation of the combined audio–mouth reliability estimate \( \xi_{12} \) is shown, as \( a_2 \) is varied from 0 to 1, which reaches a maximum value of 0.83 for an \( a_2 \) value of 0.15. So for this particular test, 0.15 and 0.85 are chosen for \( a_2 \) and \( a_1 \), respectively, i.e., the scores of module 1 are weighted more heavily.

Similarly, for Fig. 3b, the scores of module 2 are weighted more heavily. In Fig. 3c, due to the similarity of \( \xi_1 \) and \( \xi_2 \), the two modules receive approximately equal weightings. For Fig. 3d, module 2 has no contribution to the final decision. These four examples show that the weight selection procedure has the ability to adapt the weights to the reliability of each module.

Fig. 4 illustrates the fusion procedure described above, for the specific case of fusing audio–lip motion observations (\( O_{ALM} \)) with audio–lip correlation (\( O_{ALR} \)) observations. Note that the reliability estimation and weight determination occur at the score level. The described method is somewhat similar to that of [28] where the difference of log likelihood values between the hypotheses is maximised by optimising the audio and visual stream weights. The scheme differs from the method described here because it is iterative, requires a large amount of test data, operates at the frame level and is applied to speech recognition whereas the proposed scheme operates at the sentence level (i.e., the output score for an entire audio or video utterance or a test face image) and only requires a set of scores from each modules, i.e., the proposed method does not require a large amount of test data to optimise the fusion weights.

The Bi-module fusion method developed above can be employed to combine the output scores from any pair of identity verification modules. In order to carry out tri-module fusion of the audio–lip motion, audio–lip correlation, and 3D face modules, a
cascade approach is employed. First, the two audio-visual modules audio–lip motion (ALM) and audio–lip correlation (ALR) are combined, thus giving a “Audio–lip motion–correlation” score. This is shown in the first block of Fig. 5, where “Bi-module Fusion” refers to the general Bi-module fusion block illustrated in Fig. 4. The “3D face–mouth” scores (3LM) are subsequently fused with the audio–lip motion–correlation score to give the audio–3D face–mouth (“motion–correlation–depth”) scores (shown in the second block of Fig. 4) and hence we have a tri-module identity verification decision.

This multilevel fusion strategy consisting of combining two Bi-module fusion modules in cascade can take account of a noisy audio or video signal and also of any one of the three modules performing poorly, thus weighing the contribution of each module to the final decision appropriately. The advantage of this fusion method is that, being adaptive, the training of the fusion parameters is not required. Importantly, no assumption has been made about the type or level of audio or video noise that may cause a module to perform poorly. This is important for a practical audio–video system because learned noise statistics that are used to map the reliability estimate to the weighting parameter have been previously shown to vary with the type of degradation causing the train/test mismatch [4,5]. This compromises the mapping, as it must perform well for all types of noise (audio or video) and not just for one specific type of noise. Furthermore, the training of fusion parameters requires additional audio-visual data, which poses problems for the testing of existing audio-visual databases and also for practical applications, due to the small amounts of available audio-visual data. The proposed method requires no training data, and the weights are determined solely on the outputs scores from each module. We will now describe the three dimensional audio–video data corpus used, and the fusion experiments that were carried out using the proposed method.

5. 3D Audio visual data corpus

The AVOZES 3D stereovision database [17] was used for all the experiments described in this paper. AVOZES contains video recordings from 20 native speakers (10 male and 10 female) of Australian English. Video recordings were made using a calibrated stereo camera system. Video frames are stored as DV-AVI files in the NTSC format (29.97 Hz frame rate, 720 × 480 pixels resolution). Audio recordings were made using a mono microphone. Audio data are stored both in the DV-AVI files as well as in separate WAV files as 48 kHz 16 bit linear encoded samples. Module 6 of the corpus was used for training, and sentences from Module 4 were used for testing. The other modules were for calibration purposes (Module 1–3) or contain digits (Module 5). Module 4 comprises short words in a carrier phrase to cover phonemes and visemes. The sequences in this module form the core part of the AVOZES data corpus for statistical analyses of relationships between audio and video speech parameters. The phonemes and visemes in this module are in central position in CVC- or VCV-contexts to be free of any phonological or lexical restrictions. Module 4 contains several short sentences in CVC/VCV words enclosed by the carrier phrase “You grab/WORD/beer.”

Module 6 contains application-driven sequences with examples of continuous speech from each speaker. The three sequences are:

1. “Joe took father’s green shoe bench out.”
2. “Yesterday morning on my tour, I heard wolves here.”
3. “Thin hair of azure colour is pointless.”

Together with the first sentence, the second and third sentences were designed in such a way that they contain almost all phonemes and visemes of AuE (/w/ is the only phoneme missing).
To test the robustness of the proposed approach, both audio and the video test signals were degraded to provide a train/test mismatch. Ten levels of audio and video degradation were applied. This mild to adverse train/test mismatch noise levels emulates the operating scenarios encountered in a realistic operating environment. The audio models were trained on the “clean” audio speech, which was the original AVOZES audio data. Additive white Gaussian noise was applied to the clean audio at SNR levels ranging from 48 to 21 dB in decrements of 3 dB. In order to account for practical video conditions encountered in real operating scenarios, the video frame images were compressed using JPEG compression. Ten levels of JPEG QF were tested, with $QF = \{50, 28, 18, 14, 10, 8, 6, 4, 3, 2\}$, where a $QF$ of 100 represents the original uncompressed image. The variation of the mouth ROI images w.r.t. JPEG QF is shown in Fig. 6. JPEG blocking artifacts are evident at the lower QF levels.

6. Experimental results

The AVOZES database consists mostly of frontal faces and does not exhibit significant expression variations. However, some scans have slight in-depth pose variations and different expressions. Although the quality of the data is high, we used median filtering after three-dimensional reconstruction, first to remove the impulse noise, and then mean filtering was applied to smooth the facial surface. Module 6 of the database was used for training and Module 4 was used for testing. The neutral face image (1st frame of the sequence) was used for building the face template.

First, we report the performance of single mode audio, visual speech, and 3D face modules, followed by the performance of bi-modal fusion of audio–lip motion, audio–lip correlation modules, and finally the performance achieved with fusion of all three cascaded modules.

6.1. Performance of acoustic only module

For examining audio only performance, we built ten-mixture GMM speaker models trained with 34-dimensional audio MFCC features. Gender specific UBM were used as described. The three sentences from Module 6 in AVOZES were used for training and the 2 sentences from Module 4 were used for testing. The gender specific UBMs were trained using all three sentences from all the speakers from the separate male and female cohorts. All models were trained using the clean speech and tested using the various SNR levels. Fig. 7 shows how the audio-only module performs w.r.t. the audio degradation. The best EER of 2.4% was achieved at 48 dB. At 21 dB the EER dropped to worst possible EER of 50%.

6.2. Performance of 2D face only module

In this set of single mode experiments, the effect of the GMM mixtures on the performance of the four visual speech (lip motion) feature types ($f_{DCT}$, $f_{GRD}$, $f_{CTR}$), and concatenated ($f_{DCT} - f_{GRD} - f_{CTR}$) was tested initially. These tests were carried out using matched training and testing data sets, i.e., the original “clean” images. To examine whether the dynamic lip motion features, such as the $f_{GRD}$ and $f_{CTR}$ features, would perform better with a larger number of GMM mixtures, we increased the number of mixtures from one until a performance trend became apparent. For each lip feature type, a trend in the EERs with respect to the number of mixtures can be seen. The number of mixtures that maximised the visual speech features performance for each of the four feature types, is given in Table 1.

<table>
<thead>
<tr>
<th>Features</th>
<th>GMM mixtures</th>
<th>Clean</th>
<th>QF 50</th>
<th>QF 25</th>
<th>QF 18</th>
<th>QF 14</th>
<th>QF 10</th>
<th>QF 8</th>
<th>QF 6</th>
<th>QF 4</th>
<th>QF 3</th>
<th>QF 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>$f_{DCT}$</td>
<td>2</td>
<td>13.5</td>
<td>14.7</td>
<td>15.1</td>
<td>15.9</td>
<td>15.9</td>
<td>17.3</td>
<td>19.9</td>
<td>20.7</td>
<td>39.8</td>
<td>404</td>
<td>50.0</td>
</tr>
<tr>
<td>$f_{GRD}$</td>
<td>15</td>
<td>23.5</td>
<td>25.9</td>
<td>31.9</td>
<td>35.5</td>
<td>47.4</td>
<td>50.0</td>
<td>50.0</td>
<td>50.0</td>
<td>50.0</td>
<td>50.0</td>
<td>50.0</td>
</tr>
<tr>
<td>$f_{CTR}$</td>
<td>18</td>
<td>20.7</td>
<td>22.7</td>
<td>29.9</td>
<td>32.3</td>
<td>43.8</td>
<td>50.0</td>
<td>50.0</td>
<td>50.0</td>
<td>50.0</td>
<td>50.0</td>
<td>50.0</td>
</tr>
<tr>
<td>$f_{DCT} - f_{GRD} - f_{CTR}$</td>
<td>4</td>
<td>8.1</td>
<td>9.6</td>
<td>10.8</td>
<td>12.0</td>
<td>12.4</td>
<td>16.7</td>
<td>24.3</td>
<td>32.7</td>
<td>50.0</td>
<td>50.0</td>
<td>50.0</td>
</tr>
</tbody>
</table>

Fig. 6. Ten levels of JPEG compression on mouth ROI images.

Fig. 7. Effects of audio degradation on fusion performance.

Table 1
Number of Gaussian mixtures that maximises the EER performance for each of the four types of visual speech features across ten levels of JPEG Q.
The $f_{DCT}$ features performed best even with just two mixtures and decreased steadily with increasing number of states. The number of states, that maximised the EERs for the $f_{GRD}$ and $f_{CTR}$ features, were 15 and 18, respectively. The concatenated $f_{DCT} - f_{GRD} - f_{CTR}$ feature vector was modelled best using four mixtures.

For the video degradation experiments the mouth module GMMs were trained on the "clean" (uncompressed) video images and tested on the degraded video images. This provided for a mismatch between the testing and training video conditions. The tests on the degraded mismatched video data were carried out using different number of mixtures, which maximised the performance for each of the four visual feature types (as above).

This is different to all the experiments conducted for audio only mode, where ten Gaussian mixtures were always used for both training and testing. Table 1 and Fig. 8 show how different visual speech features perform w.r.t. JPEG degradation.

### 6.3. Performance of 3D face module

For this module two 3D facial features were fused with the sum rule: one for the shape modality and one for the texture modality. For the shape modality, an ICP-based point cloud algorithm was chosen. In Table 2, we provide the results of three ensembles, with TEX-GABOR being the best choice for the texture-based module. In the shape category, three different shape modules are tested, namely PC-XYZ, DI-DCT, CURV-PD, each a best representative of its own group [16]. From the performance figures in Table 2, one can see that the combination of TEX-GABOR and CURV-PD algorithms significantly outperforms the other two ensembles, and that this two have a classification performance comparable to that of different 3D facial features [16].

### 6.4. Performance of fusion of 2D–3D face features module

The face gallery (training) set, comprising three images, was formed by arbitrarily extracting the first image frame from each of the first three training sentences from AVOZES module 6. These were used to form a face template for each of the $N$ subjects.

For all the face experiments, the probe images used for testing was obtained from the module 4 sentences (again, the first frame). The gallery sets consisted of the original uncompressed images and the probe sets consisted of degraded images at the ten levels of JPEG compression. This provided for a gallery/probe mismatch. Fig. 9 and Table 3 show the performance of this module.

Different sets of experiments were similarly conducted to evaluate the performance of the audio–lip cross modal fusion features in terms of DET curves and equal error rates (EER).

<table>
<thead>
<tr>
<th>Texture module</th>
<th>Shape module</th>
<th>Performance with $E_1$, EER [%]</th>
</tr>
</thead>
<tbody>
<tr>
<td>TEX-GABOR</td>
<td>PC-XYZ</td>
<td>10.02</td>
</tr>
<tr>
<td>TEX-GABOR</td>
<td>DI-DCT</td>
<td>15.33</td>
</tr>
<tr>
<td>TEX-GABOR</td>
<td>CURV-PD</td>
<td>6.37</td>
</tr>
</tbody>
</table>

---

**Fig. 8.** Effects of video quality degradation on fusion performance.

**Fig. 9.** Fusion performance for 3D face–2D Lip motion features.
### Table 3

<table>
<thead>
<tr>
<th>JPEG QF</th>
<th>50</th>
<th>25</th>
<th>18</th>
<th>14</th>
<th>10</th>
<th>8</th>
<th>6</th>
<th>4</th>
<th>3</th>
<th>2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mouth [%]</td>
<td>14.1</td>
<td>14.9</td>
<td>15.7</td>
<td>15.7</td>
<td>17.3</td>
<td>19.8</td>
<td>20.6</td>
<td>39.5</td>
<td>50.0</td>
<td>50.0</td>
</tr>
<tr>
<td>Face [%]</td>
<td>1.2</td>
<td>1.2</td>
<td>0.4</td>
<td>0.4</td>
<td>1.2</td>
<td>1.2</td>
<td>2.0</td>
<td>8.1</td>
<td>14.1</td>
<td>25.0</td>
</tr>
<tr>
<td>Mouth-face [%]</td>
<td>0.0</td>
<td>0.8</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>1.6</td>
<td>7.3</td>
<td>12.5</td>
</tr>
</tbody>
</table>

### Table 4

<table>
<thead>
<tr>
<th>AVOZES dataset</th>
<th>JPEG QF = 50</th>
<th>Audio SNR 50 dB</th>
<th>Audio SNR 48 dB</th>
<th>Audio SNR 45 dB</th>
<th>Audio SNR 48 dB</th>
<th>Audio SNR 45 dB</th>
<th>Audio SNR 50 dB</th>
<th>Audio SNR 48 dB</th>
<th>Audio SNR 45 dB</th>
<th>Audio SNR 50 dB</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fusion of correlated f_{\text{mfcc-eiglip}} (CMF) and uncorrelated f_{\text{mfcc-eiglip}} features</td>
<td>Audio SNR 50 dB</td>
<td>Audio SNR 48 dB</td>
<td>Audio SNR 45 dB</td>
<td>Audio SNR 50 dB</td>
<td>Audio SNR 48 dB</td>
<td>Audio SNR 45 dB</td>
<td>Audio SNR 50 dB</td>
<td>Audio SNR 48 dB</td>
<td>Audio SNR 45 dB</td>
<td>Audio SNR 50 dB</td>
</tr>
<tr>
<td>f_{\text{mfcc-eiglip}}</td>
<td>4.7</td>
<td>5.18</td>
<td>8.09</td>
<td>5.81</td>
<td>6.28</td>
<td>9.74</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>f_{\text{mfcc}} * f_{\text{mfcc-eiglip}}</td>
<td>1.03</td>
<td>4.22</td>
<td>6.03</td>
<td>1.12</td>
<td>4.62</td>
<td>6.66</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>f_{\text{mfcc}} + f_{\text{mfcc-eiglip}}</td>
<td>0.68</td>
<td>0.86</td>
<td>1.29</td>
<td>0.79</td>
<td>1.17</td>
<td>1.34</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>f_{\text{mfcc}} * f_{\text{mfcc-eiglip}} + f_{\text{mfcc-eiglip}}</td>
<td>1.26</td>
<td>2.66</td>
<td>3.46</td>
<td>1.26</td>
<td>2.43</td>
<td>4.31</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>f_{\text{mfcc}} + f_{\text{mfcc-eiglip}} + f_{\text{mfcc-eiglip}}</td>
<td>1.06</td>
<td>1.85</td>
<td>2.22</td>
<td>1.23</td>
<td>2.31</td>
<td>2.46</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Table 5

<p>| Tri-module performance {ALM + ALR + 3LM} for different levels of JPEG QF and AUDIO SNRs in terms of EERs |
|-------------------------------------------------|-------------------------------------------------|-------------------------------------------------|-------------------------------------------------|-------------------------------------------------|</p>
<table>
<thead>
<tr>
<th>Audio face-mouth QF</th>
<th>db</th>
<th>48</th>
<th>45</th>
<th>42</th>
<th>39</th>
<th>36</th>
<th>33</th>
<th>30</th>
<th>27</th>
<th>24</th>
<th>21</th>
</tr>
</thead>
<tbody>
<tr>
<td>50</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>25</td>
<td>0.8</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>18</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>14</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>10</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>8</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>6</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>1.6</td>
<td>0</td>
<td>0</td>
<td>0.4</td>
<td>0</td>
<td>0</td>
<td>0.4</td>
<td>0.8</td>
<td>1.2</td>
<td>1.6</td>
<td>1.6</td>
</tr>
<tr>
<td>3</td>
<td>7.3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.8</td>
<td>1.2</td>
<td>1.6</td>
<td>2.4</td>
<td>3.6</td>
</tr>
<tr>
<td>2</td>
<td>12.5</td>
<td>0.8</td>
<td>0.8</td>
<td>0.8</td>
<td>1.2</td>
<td>1.2</td>
<td>2.4</td>
<td>2.8</td>
<td>3.6</td>
<td>4.4</td>
<td>7.3</td>
</tr>
</tbody>
</table>

6.5. Performance of fusion of acoustic-2D cross modal fusion module

As can be seen in % EER table in Table 4, the performance of ordinary features fusion of audio lip features f_{\text{mfcc-eiglip}} can be improved by cross modal analysis. The fusion of the correlated components f_{\text{mfcc-eiglip}} with mutually independent audio–lip components f_{\text{mfcc-eiglip}} results in improvement in robustness against audio SNR and video QF degradation in addition to improvement in the EER performance. Since each modality also carries mutually independent information, e.g., the texture of the lip region, possibly containing the information about the identity of a speaker, the overall performance can be enhanced with this hybrid fusion, with an optimal combination of the feature-level and the late fusion techniques combining lip, audio, and correlated audio–lip feature vectors.

6.6. Performance of fusion of tri-module fusion

For this set of experiments, we examined the performance of all the three modules, involving the fusion of audio–lip motion module (ALM module), audio–lip correlation module (ALR module) features, and the fusion of lip motion and 3D face features (3LM) in a cascaded fusion strategy shown in Fig. 5. The results for this set are shown in Table 5.

7. Discussion

As can be seen in Fig. 7, the audio module performs very well under near “clean” testing conditions, however the performance roll off w.r.t. SNR is very high. This highlights the vulnerability of a unimodal acoustic based modeling approach to mismatched train and test conditions.

For the lip motion module experiments, the fact that the DCT visual features performed best with just two Gaussian mixtures indicates that GMMs can really model the static lip shapes better, and there is not need for exploring other complex type of Gaussian models such as HMMs and embedded HMMs. Here, we have used explicit DCT based lip motion features f_{\text{GRD}} and f_{\text{CTR}} in addition to static f_{\text{DCT}} features. The best mouth module performance of 8% is surprisingly high, considering that only mouth information was employed. While the f_{\text{DCT}} – f_{\text{GRD}} – f_{\text{CTR}} concatenated features outperform the static DCT features for high QFs, f_{\text{DCT}} – f_{\text{GRD}} – f_{\text{CTR}} 9.6% versus f_{\text{DCT}} 14.3% at a QF of 50, the performance at a QF of 2 is 50%, which is similar to f_{\text{DCT}} performance. The lip motion features perform very poorly for low QF levels, both falling to around 50% at a QF of 2. The results also show that the static DCT features are more important, and are more robust than the dynamic lip motion features for the identity verification scenario.

It was expected that the 3D face module, employing features located throughout the entire face would outperform the visual speech module, which employ features extracted from just the mouth ROI. The 3D face module outperformed the mouth module at all levels of train/test mismatch. The highest face module performance was 1.2% EER, which is 15% better (relative) than the highest mouth EER accuracy. The face module also exhibits higher robustness to JPEG compression, when compared to the mouth module, with EERs less than 2%, for all test mismatch levels exceeding a QF of 4.
For the fusion of the 3D face and lip motion modules, a perfect face–lip motion EER of 0% is achieved at several levels of JPEG QF mismatch. Also, the face–lip motion mouth EERs are lower than either of the face or lip motion module EERs for all levels of JPEG QF mismatch, i.e., we have synergistic fusion. The most significant improvements are obtained for the higher levels of mismatch, for example at the lowest QF level of 2, the face–lip motion, face and lip motion EERs are 12.5%, 25%, and 50%, respectively. The performance of the face and lip motion modules both roll off suddenly at a QF of 4. This is also the case for the face–lip motion EERs, which are approximately 0% until a QF of 4 and then rise up, albeit with a lower roll off compared to either the face or lip motion modules. The improved face–lip motion performance indicates that the mouth features complement the facial features that the 3D face module uses. The improvement may be due to two factors. Firstly the 3D face module emphasises eye information and hence the mouth module is complementary, and secondly the mouth module obtains face–lip motion performance indicates that the mouth modules both roll off suddenly at a QF of 4. This is also the case for the face–lip motion EERs, which are approximately 0% until a QF of 4 and then rise up, albeit with a lower roll off compared to either the face or lip motion modules. The improved face–lip motion performance indicates that the mouth features complement the facial features that the 3D face module uses. The improvement may be due to two factors. Firstly the 3D face module emphasises eye information and hence the mouth module is complementary, and secondly the mouth module can capture the variation of the mouth ROI better over the three training video frame sequences.

The audio–lip motion EER performance also represents an improvement over the individual audio and mouth module performance at all tested levels of audio and video train/test mismatch. At the (21 dB, 2QF) operating point, the audio, mouth, and audio–mouth EERs are 50%, 50%, and 28.6%, respectively, representing a relative improvement of 49% over the mouth module.

Also, for audio–lip cross modal fusion module, the hybrid fusion involving late fusion of audio features with feature-level fusion of correlated audio–lip features based on CFA analysis \( \text{mfcc} + \text{mfcc} / \text{eigLip} \), yields a best EER performance. Further, an improvement in robustness can be observed for different audio SNRs and video QFs, for different combinations of correlated component and independent component fusion. A significant improvement in EER is achieved with correlated component hybrid fusion \( \text{mfcc} + \text{mfcc} / \text{eigLip} \) as compared to uncorrelated component hybrid fusion \( \text{mfcc} + \text{mfcc} / \text{eigLip} \). It can also be noted that all the hybrid fusion models resulted in synergistic fusion, with the EER performance better than baseline audio only and visual only EERs.

For the experiments involving tri-module fusion \{ALM + ALR + 3LM\}, perfect EERs of 0% were achieved at the majority of operating points. The tri-module fusion attains a significant increase in robustness to both audio and video degradations. This is evident from the flatness of the EERS in Table 5.

8. Conclusions

In this paper, a robust spatio-temporal face modelling approach involving a multilevel fusion strategy based on cascaded multimodal fusion of audio, lip motion, correlation, and 3D features is proposed for biometric speaker identity verification scenario. The approach uses the biometric information from three modules, namely audio, visual speech, and 3D face data, extracts optimal motion, correlation and depth (3D) features, and performs a hybrid cascaded fusion, adapting to the local performance of each module, and taking into account the output-score based reliability estimates of each of the modules. These results as a whole are important for remote authentication applications, where bandwidth is limited and uncontrolled acoustic noise is probable, such as video telephony and online authentication systems. The performance evaluation for liveness verification scenario, which allows detection of fraudulent replay attacks on biometric security system are the objectives of further research.

References


