Audio-Visual Identity Verification: An Introductory Overview

Bouchra Abboud, Hervé Bredin, Guido Aversano, and Gérard Chollet

CNRS-LTCI, GET-ENST, 46 rue Barrault, 75013 PARIS, FRANCE,
bouchra.abboud@facing-it.net - {bredin, aversano, chollet}@tsi.enst.fr

Abstract. Verification of identity is commonly achieved by looking at the face of a person and listening to his (her) speech. Automatic means of achieving this verification has been studied for several decades. Indeed, a talking face offers many features to achieve a robust verification of identity. The current deployment of videophones drives new opportunities for a secured access to remote servers (banking, certification, call centers, etc.). The synchrony of the speech signal and lip movements is a necessary condition to check that the observed talking face has not been manipulated and/or synthesized. This overview addresses face, speaker and talking face verification, as well as face and voice transformation techniques. It is demonstrated that a dedicated impostor needs limited information from a client to fool state of the art audio-visual identity verification systems.

1 Introduction

Identity verification based on talking face biometrics is getting more and more attention. As a matter of fact, a talking face offers many features to achieve a robust identity verification: it includes speaker verification, face recognition and their multimodal combination. Moreover, whereas iris or fingerprint biometrics might appear intrusive and need user collaboration, a talking face identity verification system is not intrusive and can even be achieved without the user noticing it.

Though it can be very robust thanks to the complementarity of speaker and face recognition, these two modalities also share a common weakness: an impostor can easily record the voice or photograph the face or his/her target without him/her noticing it ; and thus fool a talking face system with very little effort. Moreover, higher effort impostors might perform both voice conversion and face animation in order to perform impersonation that is even more difficult to detect or to mask his/her identity (see *Voice Disguise and Automatic Detection: Review and Perspectives* by Perrot et al. in this volume). We will show that explicit talking face modeling (i.e. the coupled modeling of acoustic and visual synchronous features) is an effective way to overcome these weaknesses.

The remaining of this chapter is organized as follows. After a short review of the most prominent methods used for speaker verification and face recognition, low and high forgery scenarios are described, including simple replay attacks,

voice conversion and face animation. Finally, the particular task of replay attacks detection is addressed, based on the detection of a lack of synchrony between voice and lip motion.

2 Audio-Visual Identity Verification

2.1 Speaker Verification

Speech is a biometric modality that may be used to verify the identity of a speaker. The speech signal represents the amplitude of an audio waveform as captured by a microphone. To process this signal a feature extraction module calculates relevant feature vectors on a signal window that is shifted at a regular rate. In order to verify the identity of the claimed speaker a stochastic model for the speech generated by the speaker is generally constructed. New utterance feature vectors are generally matched against the claimed speaker model and against a general model of speech that may be uttered by any speaker called the world model. The most likely model identifies if the claimed speaker has uttered the signal or not. In text independent speaker recognition, the model should not reflect a specific speech structure, i.e. a specific sequence of words. Therefore in state-of-the-art systems, Gaussian Mixture Models (GMM) are used as stochastic models [1].

Given a feature vector $\mathbf{x}$, the GMM defines its probability distribution function as in (1).

$$\sum_{i=1}^{N} w_i \frac{1}{\sqrt{(2\pi)^d \|\Gamma_i\|}} \exp\left(-\frac{1}{2}(\mathbf{x}-\mu_i)^T \Gamma_i^{-1}(\mathbf{x}-\mu_i)\right) \tag{1}$$

This distribution can be seen as the realizations of two successive processes. In the first process, the mixture component is selected and based on the selected component the corresponding Gaussian distribution defines the realization of the feature vector. The GMM model is defined by the set of parameters $\lambda = (\{w_i\}, \{\mu_i\}, \{\Gamma_i\})$. To estimate the GMM parameters, speech signals are generally collected. The unique observation of the feature vectors provides incomplete data insufficient to allow analytic estimation, following the maximum likelihood criterion, of the model parameters, i.e. the Gaussian distributions weights, mean vectors and covariance matrices. The Estimation Maximization (EM) algorithm offers a solution to the problem of incomplete data [2]. The EM algorithm is an iterative algorithm, an iteration being formed of two phases: the Estimation (E) phase and the Maximization (M) phase. In the E phase the likelihood function of the complete data given the previous iteration model parameters is estimated. In the M phase new values of the model parameters are determined by maximizing the estimated likelihood. The EM algorithm ensures that the likelihood on the training data does not decrease with the iterations and therefore converges towards a local optimum. This local optimum depends on the initial values given to the model parameters before training. Thus, the

initialization of the model parameters is a crucial step. The LBG algorithm is used to initialize the model parameters.

The direct estimation of the GMM parameters using the EM algorithm requires a large amount of speech feature vectors. This causes no problem for the world model where several minutes from several speakers may be collected for this purpose. For the speaker model, this would constrain the speaker to talk for a long duration and may not be acceptable. To overcome this, speaker adaptation techniques may be used [3], such as Bayesian adaptation, maximum likelihood linear regression (MLLR), and the unified adaptation technique defined in [4]. Using the adaptation techniques few minutes of speech become sufficient to determine the speaker model parameters.

During recognition, feature vectors are extracted from a speech utterance. The log likelihood ratio between the speaker and world models is computed and compared to a threshold. This allows to verify the identity of the claimed speaker.

2.2 Face Verification

Face recognition is divided into two major areas: face identification and face verification. On the one hand, face verification is concerned with validating a claimed identity based on a frontal and/or profile image or a video sequence of the claimant's face, and either accepting or rejecting the identity claim. On the other hand, the goal of face identification is to identify a person based on the image or video of his/her face. This face has to be compared with all registered persons and hence it is desirable to represent faces in a compact yet precise manner. Many techniques exist to perform this task, some of which are reviewed below.

Face Representation. Traditionally two major classes of techniques exist for face representation. On the one hand, geometrical feature extraction relies on parameters of distinctive features such as eyes, mouth and nose. On the other hand, in appearance-based approaches, a face is represented as an array of intensity values suitably preprocessed; the array is then compared with a face template using an appropriate metric. The performances of both representation techniques in face recognition are compared in [5]. Our work on faces belongs to appearance-based approaches.

Face Recognition. Once proper face representation is accomplished, the next step consists in determining which class the represented face belongs to. In this context several approaches have been proposed.
In template matching, the extracted face information is compared with the precomputed templates of each class [6]. The degree of similarity is measured either with the Euclidian distance or with the Mahalanobis distance in eigenspace and fisherspace [7, 8]. A probabilistic similarity measure is also used in [9].
However, in distance-based classification, it is assumed that the prototypes are

representative of query images under various conditions and the recognition performance depends largely on the representational capacity of the training set. Therefore, location and scale of query images are usually normalized before they are compared to the templates. Nevertheless, changes in illumination and rotation are difficult to compensate by normalization. In this context, the nearest feature line method proposed in [10] aims at expanding the representational capacity of available feature points in order to account for new conditions not represented in the training set. The method interpolates the available prototype images to build appropriate linear combinations that represent the variations in illumination and viewing angle. The decision is made based on the minimum distance between the query and each interpolated line.

Support Vector Machine (SVM) is an effective method for pattern recognition that finds the hyperplane separating the largest possible fraction of points of the same class on the same side while maximizing the distance from either classes to the hyperplane [11]. SVM is a binary classifier, and two strategies exist for solving q-class problems. The one-versus-all strategy involves the training of q SVMs, each separating a given class from the rest of the training set, whereas pairwise classifiers involve the training of a different SVM for separating each pair of classes. The one-versus-all strategy was successfully used for facial expression recognition [12] as well as face recognition [13].

Multilayer Perceptron (MLP) neural network (NN) is a good classification tool. It searches for an acceptable local minimum in the NN weight space in order to achieve minimal error. Weights are adjusted using back-propagation which is a gradient descent supervised training procedure. During the training procedure MLP builds separating hypersurfaces in the input space. After training MLP can successfully apply acquired skills to previously unseen samples. Backpropagation trained Neural Networks were used for facial expression recognition in [14].
Hidden Markov Model (HMM) is an extension of the theory of Markov chains where the observation of a certain output is a probabilistic function of the state. Identification is achieved by selecting the HMM which obtains the highest likelihood [15]. HMM computations converge quickly making them practical for real time processing.

Finally, recent papers investigate the use of video sequences in order to perform face recognition. For instance, in [16], face features are extracted from every frame of the video, and gaussian mixture modeling is used (as in the speaker verification task). In the GMM space, an additional step is performed: one SVM is trained per client to achieve better discrimination between clients.

3 The Forgery Issue in Biometrics

Many databases are available to the research community to help evaluate multimodal biometric verification algorithms, such as BANCA [17], BT-DAVID [18], XM2VTS [19] and BIOMET [20]. Different protocols have been defined for evaluating biometric systems on each of these databases, but they share the assumption that impostor attacks are zero-effort attacks. For example, in the particular

framework of the BANCA database, each subject records one client access and one impostor access per session. However, the only difference between the two is the particular message that the client utters—their name and address in the first case; the target's name and address in the second. Thus the impersonation takes place without any knowledge of the target's face, age, and voice. These zero-effort impostor attacks are unrealistic—only a fool would attempt to imitate a person without knowing anything about them. In this work we adopt more realistic scenarios in which the impostor has more information about the target.

3.1 Replay Attacks

A major drawback of using the talking-face modality for identity verification is that an impostor can easily obtain a sample of any client's audiovisual identity. Contrast this with iris recognition: it is quite difficult to acquire a sample of another person's iris. But numerous small devices allow an impostor to take a picture of the target's face without being noticed, and some mobile phones are even able to record movies. Of course, it is even easier to acquire a recording of the target's voice. Therefore, protocols to evaluate audiovisual identity verification systems should recognize this fact, for example by adding replay attacks to their repertoire of envisaged impostor accesses [21].

3.2 Voice Conversion

Forgery attacks against a speaker verification system, where the voice characteristics of the impostor is modified in such a way that it resembles the voice of the client, are investigated in this section. The choice of the transformation technique is made according to the available quantity of client voice data.

If only a limited amount of client data is available for training (as in the case of BANCA protocol [17]), a spectral conversion technique should be adopted, which give some interesting results according to previous studies [22]. Consider a sequence of spectral vectors pronounced by the impostor, $X = [x_1, x_2, \ldots, x_n]$, and a sequence composed by the same words, pronounced by the client, $Y = [y_1, y_2, \ldots, y_n]$. A spectral transformation can be performed by finding the conversion function F that minimizes the mean square error: $\epsilon_{\text{mse}} = \mathbb{E}[||y - F(x)||^2]$, where $\mathbb{E}$ is the expectation. This conversion method requires preliminary word-level segmentation of the training sentences.

If more client data are available (e.g. approximately 1 hour of client's speech), we can use a voice encoder based on the Automatic Language Independent Speech Processing approach (ALISP) [23]. This method is described in Sec.2.2 of the chapter untitled *Voice Disguise and Automatic Detection: Review and Program* by Perrot et al.

3.3 Face Transformation

Natural talking faces synthesis is a very challenging task, since a synthetic face has to be photo-realistic and represent subtle texture and shape variations that

are vital to talking faces representation and recognition, in order to be considered natural.

Many modeling techniques exist which achieve various degrees of realism and flexibility. The first class of techniques uses 3D meshes to model the face shape [24, 25]. To obtain natural appearance a 3D scan image of the subject face is texture-mapped on the 3D parameterized deformable model.

An alternative approach is based on morphing between 2D images. This technique produces photo-realistic images of new shapes by performing interpolation between previously seen shapes and is successfully combined with geometric 3D transformations to create realistic facial models from photos and construct smooth transitions between different facial expressions [26]. Using the same technique, multi-dimensional deformable models [27] can generate intermediate video-realistic mouth movements of a talking face from a small set of manually selected mouth samples. Morphing is also used in the context of the *video-rewrite* to change the identity of a talking face [28].

3.4 Talking-Face Animation

In this work we propose to use an appearance-based face tracker allowing to extract from each frame of a video sequence a set of feature points describing the face shape. These feature points are tracked from frame to frame throughout the entire sequence [29] and their motion is injected into any target image allowing to simulate a lip movement similar to the tracked sequence.

Face tracking. It has already been shown that the active appearance model [30] is a powerful tool for object synthesis and tracking. It uses Principal Component Analysis (PCA) to model both shape and texture variations seen in a training set of visual objects. After computing the mean shape $\bar{\mathbf{s}}$ and aligning all shapes from the training set by means of a Procrustes analysis, the statistical shape model is given by (2)

$$\mathbf{s}_i = \bar{\mathbf{s}} + \Phi_s \mathbf{b}_{si} \tag{2}$$

where $\mathbf{s}_i$ is the synthesized shape, Φ_s is a truncated matrix describing the principal modes of shape variations in the training set and $\mathbf{b}_{si}$ is a vector that controls the synthesized shape.

It is then possible to warp textures from the training set of faces onto the mean shape $\bar{\mathbf{s}}$ in order to obtain shape-free textures. Similarly, after computing the mean shape-free texture $\bar{\mathbf{t}}$ and normalizing all textures from the training set relatively to $\bar{\mathbf{t}}$ by scaling and offset of the luminance values, the statistical texture model is given by (3)

$$\mathbf{t}_i = \bar{\mathbf{t}} + \Phi_t \mathbf{b}_{ti} \tag{3}$$

where $\mathbf{t}_i$ is the synthesized shape-free texture, Φ_t is a truncated matrix describing the principal modes of texture variations in the training set and $\mathbf{b}_{ti}$ is a vector that controls the synthesized shape-free texture.

By combining the training shape and texture vectors $\mathbf{b}_{si}$ and $\mathbf{b}_{ti}$ and applying further PCA the statistical appearance model is given by (4) and (5)

$$\mathbf{s}_i = \bar{\mathbf{s}} + Q_s \mathbf{c}_i \tag{4}$$

$$\mathbf{t}_i = \bar{\mathbf{t}} + Q_t \mathbf{c}_i \tag{5}$$

where Q_s and Q_t are truncated matrices describing the principal modes of combined appearance variations in the training set, and $\mathbf{c}_i$ is a vector of appearance parameters simultaneously controlling both shape and texture.

Given the parameter vector $\mathbf{c}_i$, the corresponding shape $\mathbf{s}_i$ and shape-free texture $\mathbf{t}_i$ can be computed respectively using (4) and (5). The reconstructed shape-free texture is then warped onto the reconstructed shape in order to obtain the full appearance. Displacing each modes of the mean appearance vector $\bar{\mathbf{c}}$ changes both the texture and shape of the coded synthetic faces.

Furthermore, in order to allow pose displacement of the model, it is necessary to add to the appearance parameter vector $\mathbf{c}_i$ a pose parameter vector $\mathbf{p}_i$ allowing control of scale, orientation and position of the synthesized faces.

While a couple of appearance parameter vector $\mathbf{c}$ and pose parameter vector $\mathbf{p}$ represents a face, the active appearance model can automatically adjust those parameters to a target face by minimizing a residual image $\mathbf{r}(\mathbf{c}, \mathbf{p})$ which is the texture difference between the synthesized faces and the corresponding mask of the image it covers as shown in (6) and (7).

In the following, the appearance and pose parameters obtained by this optimization procedure will be denoted respectively as $\mathbf{c}_{op}$ and $\mathbf{p}_{op}$.

$$\mathbf{c}_{op} = \arg_{min} |r[(\mathbf{c} + \delta\mathbf{c}), \mathbf{p}]|^2 \tag{6}$$

$$\mathbf{p}_{op} = \arg_{min} |r[\mathbf{c}, (\mathbf{p} + \delta\mathbf{p})]|^2 \tag{7}$$

For this purpose, a set of training residual images are computed by displacing the appearance and pose parameters within allowable limits. These residuals are then used to compute matrices $\mathbf{R}_a$ and $\mathbf{R}_t$ establishing the linear relationships (8) and (9)

$$\delta(\mathbf{c}) = -\mathbf{R}_a \, \mathbf{r}(\mathbf{c}, \mathbf{p}) \tag{8}$$

$$\delta(\mathbf{p}) = -\mathbf{R}_t \, \mathbf{r}(\mathbf{c}, \mathbf{p}) \tag{9}$$

between the parameter displacements and the corresponding residuals, so as to minimize $|\mathbf{r}((\mathbf{c}, \mathbf{p}) + \delta(\mathbf{c}, \mathbf{p}))|^2$.

A first order Taylor development gives the following solution (10) and (11)

$$\mathbf{R}_a = \left(\frac{\partial \mathbf{r}}{\partial \mathbf{c}}^T \frac{\partial \mathbf{r}}{\partial \mathbf{c}} \right)^{-1} \frac{\partial \mathbf{r}}{\partial \mathbf{c}}^T \tag{10}$$

$$\mathbf{R}_t = \left(\frac{\partial \mathbf{r}}{\partial \mathbf{p}}^T \frac{\partial \mathbf{r}}{\partial \mathbf{p}} \right)^{-1} \frac{\partial \mathbf{r}}{\partial \mathbf{p}}^T \tag{11}$$

These linear relationships are then used to determine the optimal appearance and pose vectors $\mathbf{c}_{op}$ and $\mathbf{p}_{op}$ using a gradient descent algorithm [31].

Hence adapting the active appearance model to each frame of a video sequence showing a speaking face allows to track the facial movements of the face as shown on Fig. 1. The experiments are conducted on the BANCA database

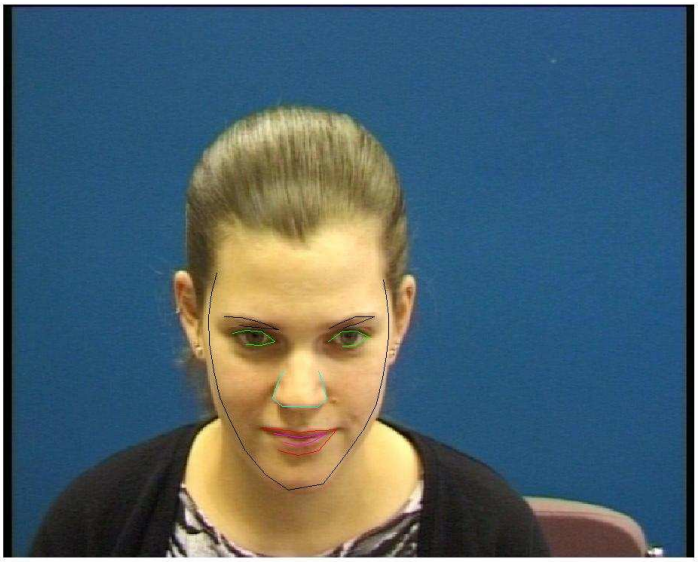

Fig. 1. Face tracking through consecutive adaptation of AAM to each frame.

which contains video recordings of different speaking faces. Evaluation is conducted according to the MC evaluation protocol [32]. This database was designed in order to test multi-modal (face and voice) identity verification with various acquisition devices. For 4 different languages (English, French, Italian and Spanish), video and speech data were collected for 52 subjects on 12 different occasions. During each recording, the subject was prompted to say a random 12 digits number, their name, address and date of birth.

The consecutive frames were extracted from each video sequence of the BANCA database. An appearance model is first trained using the first 5 images of the training set, which corresponds to the world model according to the protocol, and then used to automatically detect feature points on the next 5 images. The model is subsequently rebuilt using the whole 10 annotated images and so on to annotate the whole training set in a bootstrapping mode. The obtained model is then used to automatically annotate the client verification data.

This procedure allows hence to perform automatic face tracking on the client verification sequences.

Face Animation. Lip motion is defined by the position of the MPEG-4 compatible feature points on each frame of the tested sequence. A set of 18 features points were selected at key positions on the outer and inner lip contours as shown on Fig. 2. This motion can be injected to any target image showing an unknown

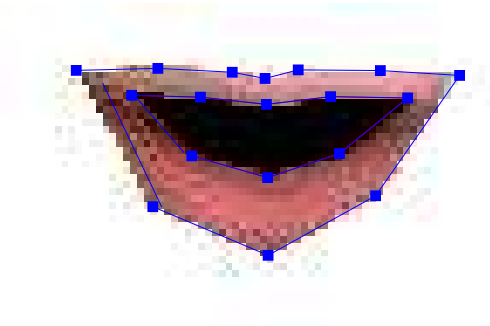

Fig. 2. MPEG-4 compatible feature points located at the inner and outer contours of the training lips.

face using the following procedure.

First the lip pixels on the target image are detected using the lip localization method described in Sec. 3.4, as shown on Fig. 1. Then, starting from this rough position, an appearance model is initialized and adapted to the target lip using the gradient descent algorithm. This procedure allows to automatically place the feature points at the correct positions of the target lips as shown on Fig. 3. An artificial motion is then obtained by displacing these feature points to match each frame of the driving sequence. A Delaunay triangulation coupled with a piecewise affine transform is used to interpolate pixels color values. An example of lip motion cloning of a driving sequence on an unknown target face is shown on Fig. 4.

4 Replay Attacks Detection

The main weakness of a biometric system based on the fusion of speaker verification and face recognition stays in the fact that it is easily fooled by replay attacks (as described in Sec. 3). The solution that we propose is to detect a lack of synchrony between voice and lip motion that could result from this kind of replay attacks.

In this section, we overview the most promising methods of the literature allowing to measure the degree of synchrony between audio and visual speech [33].

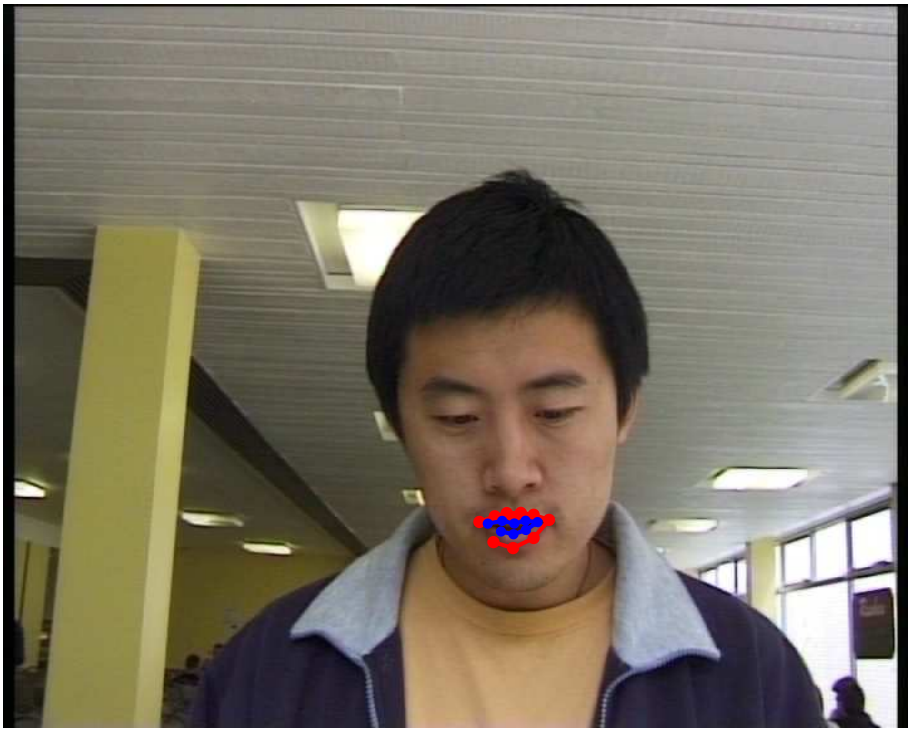

Fig. 3. Appearance model adaptation and automatic feature point placement on the target face.

4.1 Audio-Visual Features

Most of the audio-visual features used in the relatively new field dealing with audio-visual speech synchrony are shared with the audio-visual speech recognition domain [34].

Audio Speech. Acoustic speech parameterization is classically performed on overlapping sliding window of the original audio signal.

In [35], the authors extract the average acoustic energy on the current window as their one-dimensional audio feature, whereas [36] uses the periodogram. The use of classical Mel-Frequency Cepstral Coefficients (MFCC) is very frequent [37–41]. Linear-Predictive Coding (LPC), and its derivation Line Spectral Frequencies (LSF) [42] have also been widely investigated. The latter are often preferred because shown to be strongly related to the vocal tract geometry [43].

A comparison of these different acoustic speech features is performed in [37] in the framework of the *FaceSync* linear operator. To summarize, in their specific framework, the authors conclude that MFCC, LSF and LPC parameterizations

lead to a stronger correlation with the visual speech than spectrogram and raw energy features.

Visual Speech. Raw pixels are the visual equivalent of the audio raw energy. In [35] and [40], the intensity of gray-level pixels is used as is. In [21], their sum over the whole region of interest is computed, leading to a one-dimensional feature. Holistic methods consider and process the region of interest as a whole source of information. In [39], a two-dimensional discrete cosine transform (DCT) is applied on the region of interest. In [44], the authors perform a projection of the region of interest on vectors resulting from a principal component analysis (PCA): they call the principal components eigenlips. Lip-shape methods consider and process lips as a deformable object from which geometrical features can be derived, such as height, width openness of the mouth, position of lip corners, etc. Mouth width, mouth height and lip protrusion are computed in [45]. In [46, 47], a deformable template composed of several polynomial curves follows the lip contours: it allows the computation of the mouth width, height and area. In [41], the lip shape is summarized with a one-dimensional feature: the ratio of lip height and lip width.

Audio-Visual Subspaces. Once these features are computed, transformation are performed in order to reduce dimensionality and keep only the dimensions that are meaningful for the specific task of measuring the degree of synchrony in audio-visual speech. Thus, in [48], Principal Component Analysis (PCA) is used in order to reduce the dimensionality of a joint audiovisual space (in which audio speech features and visual speech features are concatenated), while keeping the characteristics that contribute most to its variance. In [49], Independent Component Analysis (ICA) is applied on an audiovisual recording of a piano session: the camera frames a close-up on the keyboard when the microphone is recording the music. ICA allows to clearly find a correspondence between the audio and visual note. However, to our knowledge, ICA has never been used as a transformation of the audiovisual speech feature space. Canonical Correlation Analysis (CANCOR) is a statistical analysis allowing to jointly transform the audio and visual feature spaces while maximizing the audiovisual cross-correlation. Given two synchronized random variables, the *FaceSync* algorithm presented in [37] uses CANCOR to find canonic correlation matrices that whiten them under the constraint of making their cross-correlation diagonal and maximally compact. Co-Inertia Analysis (CoIA) is quite similar to CANCOR. However, while CANCOR is based on the maximization of the correlation between audio and visual features, CoIA relies on the maximization of their covariance [50, 46].

4.2 Measures

Once audio-visual speech features are extracted, the next step is to measure their degree of correspondence. The following paragraphs overview what is proposed in the literature.

Correlation. Let X and Y be two independent random variables which are normally distributed.

Assuming a linear relationship between X and Y, the square of their Pearsons product-moment coefficient $R(X,Y)$ (defined in equation 12) denotes the portion of total variance of X that can be explained by a linear transformation of Y (and reciprocally, since it is a symmetrical measure).

$$R(X,Y) = \frac{\mathrm{cov}\,(X,Y)}{\sigma_X \sigma_Y} \tag{12}$$

In [35], the authors compute the Pearsons product-moment coefficient between the average acoustic energy X and the value Y of the pixels of the video to determine which area of the video is more correlated with the audio. This allows to decide which of two people appearing in a video is talking.

In information theory, the mutual information $MI(X,Y)$ of two random variables X and Y is a quantity that measures the mutual dependence of the two variables. In the case of X and Y are discrete random variables, it is defined as in equation (13).

$$MI(X,Y) = \sum_{x \in X} \sum_{y \in Y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)} \tag{13}$$

It is non-negative ($MI(X,Y) \geq 0$) and symmetrical ($MI(X,Y) = MI(Y,X)$). One can demonstrate that X and Y are independent if and only if $MI(X,Y) = 0$. The mutual information can also be linked to the concept of entropy H in information theory as shown in equation 15:

$$MI(X,Y) = H(X) - H(X|Y) \tag{14}$$

$$MI(X,Y) = H(X) + H(Y) - H(X,Y) \tag{15}$$

In [35, 51, 39, 40], the mutual information is used to locate the pixels in the video which are most likely to correspond to the audio signal: the face of the person who is speaking clearly corresponds to these pixels.

Joint Audio-Visual Models. Let consider two discrete random variables $X = \{x_t, t \in \mathbb{N}\}$ and $Y = \{y_t, t \in \mathbb{N}\}$ of dimension d_X and d_Y respectively. One can define the discrete random variable $Z = \{z_t, t \in \mathbb{N}\}$ of dimension d_Z where z_t is the concatenation of the two samples x_t and y_t, such as $z_t = [x_t, y_t]$ and $d_Z = d_X + d_Y$.

Given a sample z, the Gaussian Mixture Model (GMM) λ defines its probability distribution function as in 16.

$$p(z|\lambda) = \sum_{i=1}^{N} w_i \mathcal{N}(z; \mu_i, \Gamma_i) \tag{16}$$

where $\mathcal{N}(\bullet; \mu, \Gamma)$ is the normal distribution of mean μ and covariance matrix Γ. $\lambda = \{w_i, \mu_i, \Gamma_i\}_{i \in [1,N]}$ are parameters describing the joint distribution of X and

Y. Using a training set of synchronized samples x_t and y_t concatenated into joint samples z_t, the Expectation-Maximization algorithm (EM) allows the estimation of λ. Given two sequences of test $X = \{x_t, t \in [1,T]\}$ and $Y = \{y_t, t \in [1,T]\}$, a measure of their correspondence $C_\lambda(X,Y)$ can be computed as in (17).

$$C_\lambda(X,Y) = \frac{1}{T}\sum_{t=1}^{T} p([x_t, y_t]|\lambda) \tag{17}$$

The authors of [39] propose to model audio-visual speech with Hidden Markov Models (HMMs). Two speech recognizers are trained: one classical audio only recognizer [52], and an audiovisual speech recognizer as described in [34]. Given a sequence of audiovisual samples ($[x_t, y_t], t \in [1,T]$), the audio only system gives a word hypothesis W. Then, using the HMM of the audiovisual system, what the authors call a measure of plausibility $P(X,Y)$ is computed as follows:

$$P(X,Y) = p([x_1, y_1]...[x_T, y_T]|W) \tag{18}$$

An asynchronous hidden Markov model (AHMM) for audio-visual speech recognition is proposed in [53]. It assumes that there is always an audio observation x_t and sometimes a visual observation y_s at time t. It intrinsically models the difference of sample rates between audio and visual speech, by introducing the probability that the system emits the next visual observation y_s at time t. AHMM appears to outperform HMM in the task of audio-visual speech recognition [53] while naturally resolving the problem of different audio and visual sample rates.

The use of neural networks (NN) is investigated in [38]. Given a training set of both synchronized and not synchronized audio and visual speech features, a neural network with one hidden layer is trained to output 1 when the audiovisual input features are synchronized and 0 when they are not. Moreover, the authors propose to use an input layer at time t consisting of $[X_{t-N_X}, \ldots, X_t, \ldots, X_{t+N_X}]$ and $[Y_{t-N_Y}, \ldots, Y_t, \ldots, Y_{t+N_Y}]$ (instead of X_t and Y_t), choosing N_X and N_Y such as about 200 ms of temporal context is given as an input.

5 Conclusion

Biometric identity verification is usually used to protect the access to sensitive information or locations, which are –by definition– prone to be attacked by malevolent people. Therefore, it is very important to investigate the possible attacks that could threaten such a system. In the case of audio-visual identity verification based on talking faces, we have shown that it is possible to increase error rates by transforming the voice of the impostor so that it resembles the voice of his/her target. Moreover, an algorithm allowing to automatically animate the face of a person in order to reproduce the lip motion of an impostor was described. These high-effort forgeries are a great threat for talking face-based identity verification algorithms. However, much simpler attacks (yet very

efficient if the algorithm was not originally designed to take them into account) can also be used by impostors. This is the case of replay attacks, which can be detected by measuring a possible lack of synchrony between voice and lip motion.

Acknowledgments

This work was partially supported by the European Commission through our participation to the SecurePhone project (http://www.secure-phone.info/) and the BioSecure Network of Excellence (http://www.biosecure.info/)

References

1. Reynolds, D.A., Quatieri, T.F., Dunn, R.B.: Speaker Verification using Adapted Gaussian Mixture Models. Digital Signal Processing **10** (2000) 19 – 41
2. Dempster, A., Laird, N., Rubin, D.: Maximum Likelihood from Incomplete Data via the EM Algorithm. J. of Royal Statistical Society **39**(1) (1977) 1 – 22
3. Blouet, R., Mokbel, C., Mokbel, H., Sanchez, E., Chollet, G.: BECARS: a Free Software for Speaker Verification. In: ODYSSEY 2004. (2004) 145 – 148
4. Mokbel, C.: Online Adaptation of HMMs to Real-Life Conditions: A Unified Framework. In: IEEE Transactions on Speech and Audio Processing. Volume 9. (2001) 342 – 357
5. Brunelli, R., Poggio, T.: Face recognition: Features versus templates. IEEE Trans. on Pattern Analysis and Machine Intelligence **15**(10) (1993) 1042–1052
6. Wiskott, L., Fellous, J.M., Krüger, N., von der Malsburg, C.: Face recognition by elastic bunch graph matching. In: Intl. Conference on Computer Analysis of Images and Patterns. Number 1296, Heidelberg, Springer-Verlag (1997) 456–463
7. Abboud, B., Davoine, F., Dang, M.: Expressive face recognition and synthesis. In: IEEE CVPR workshop on Computer Vision and Pattern Recognition for Human Computer Interaction, Madison, U.S.A. (2003)
8. Turk, M., Pentland, A.: Eigenfaces for recognition. Journal of Cognitive Neuroscience **3**(1) (1991) 71–86
9. Moghaddam, B., Pentland, A.: Beyond euclidean eigenspaces: Bayesian matching for visual recognition. In: Face Recognition: From Theories to Applications, Berlin, Springer-Verlag (1998)
10. Li, S., Lu, J.: Face recognition using the nearest feature line method. IEEE Transactions on Neural Networks **10** (1999) 439–443
11. Vapnik, V. In: Statistical Learning Theory. Wiley (1998)
12. Bartlett, M.S., Littlewort, G., Fasel, I., Movellan, J.R.: Real time face detection and facial expression recognition: Development and applications to human computer interaction. In: IEEE CVPR workshop on Computer Vision and Pattern Recognition for Human Computer Interaction, Madison, U.S.A. (2003)
13. Heisele, B., Ho, P., J.Wu, Poggio, T.: Face recognition: Component-based versus global approaches. In: Computer Vision and Image Understanding. Volume 91. (2003) 6–21
14. Padgett, C., Cottrell, G., Adolphs, R.: Categorical perception in facial emotion classification. In: Proceedings of the Eighteenth Annual Cognitive Science Conference., San Diego, CA (1996) 249–253

15. Lien, J., Zlochower, A., Cohn, J., Li, C., Kanade, T.: Automatically recognizing facial expressions in the spatio temporal domain. In: Proceedings of the Workshop on Perceptual User Interfaces, Alberta, Canada (1997)
16. Bredin, H., Dehak, N., Chollet, G.: GMM-based SVM for Face Recognition. International Conference on Pattern Recognition (2006)
17. Bailly-Baillière, E., Bengio, S., Bimbot, F., Hamouz, M., Kittler, J., Mariéthoz, J., Matas, J., Messer, K., Popovici, V., Porée, F., Ruiz, B., Thiran, J.P.: The BANCA Database and Evaluation Protocol. In: Lecture Notes in Computer Science. Volume 2688. (2003) 625 – 638
18. BT-DAVID: (http://eegalilee.swan.ac.uk/)
19. Messer, K., Matas, J., Kittler, J., Luettin, J., Maitre, G.: XM2VTSDB: The Extended M2VTS Database. Audio- and Video-Based Biometric Person Authentication (1999) 72 – 77
20. Garcia-Salicetti, S., Beumier, C., Chollet, G., Dorizzi, B., Jardins, J.L., Lunter, J., Ni, Y., Petrovska-Delacretaz, D.: BIOMET: a Multimodal Person Authentication Database including Face, Voice, Fingerprint, Hand and Signature Modalities. Audio- and Video-Based Biometric Person Authentication (2003) 845 – 853
21. Bredin, H., Miguel, A., Witten, I.H., Chollet, G.: Detecting Replay Attacks in Audiovisual Identity Verification. In: IEEE International Conference on Acoustics, Speech, and Signal Processing. (2006)
22. Stylianou, Y., Cappé, O., Moulines, E.: Statistical Methods for Voice Quality Transformation. European Conference on Speech Communication and Technology (1995)
23. Perrot, P., Aversano, G., Chollet, G., Charbit, M.: Voice Forgery Using ALISP: Indexation in a Client Memory. In: ICASSP 2005. (2005)
24. Romdhani, S., Vetter, T.: Efficient, robust and accurate fitting of a 3D morphable model. In: IEEE Intl. Conference on Computer Vision, Nice, France (2003)
25. Terzopoulos, D., Waters, K.: Analysis and synthesis of facial image sequences using physical and anatomical models. IEEE Trans. on Pattern Analysis and Machine Intelligence **15**(6) (1993) 569–579
26. Pighin, F., Hecker, J., Lischinski, D., Szeliski, R., Salesin, D.: Synthesizing realistic facial expressions from photographs. In: Siggraph proceedings. (1998) 75–84
27. Ezzat, T., Geiger, G., Poggio, T.: Trainable videorealistic speech animation. In: ACM Siggraph, San Antonio, Texas (2002)
28. Bregler, C., Covel, M., Slaney, M.: Video rewrite: Driving visual speech with audio. In: Siggraph proceedings. (1997) 353–360
29. Ahlberg, J.: An active model for facial feature tracking. EURASIP Journal on applied signal processing **6** (2002) 566–571
30. Abboud, B., Davoine, F., Dang, M.: Facial expression recognition and synthesis based on an appearance model. Signal Processing: Image Communication **10**(8) (2004) 723–740
31. Cootes, T., Edwards, G., Taylor, C.: Active appearance models. IEEE Trans. on Pattern Analysis and Machine Intelligence **23**(6) (2001) 681–685
32. Bailly-Bailliere, E., Bengio, S., Bimbot, F., Hamouz, M., Kittler, J., Mariethoz, J., Matas, J., Messer, K., Popovici, V., Pore, F., Ruiz, B., Thiran, J.P.: The BANCA database and evaluation protocol. In: 4th International Conference on Audio- and Video-Based Biometric Person Authentication, AVBPA, Springer-Verlag (2003)
33. Bredin, H., Chollet, G.: Measuring Audio and Visual Speech Synchrony: Methods and Applications. In: International Conference on Visual Information Engineering. (2006)

34. Potamianos, G., Neti, C., Luettin, J., Matthews, I.: 10. In: Audio-Visual Automatic Speech Recognition: An Overview. MIT Press (2004)
35. Hershey, J., Movellan, J.: Audio-Vision: Using Audio-Visual Synchrony to Locate Sounds. Neural Information Processing Systems (1999)
36. Fisher, J.W., Darell, T.: Speaker Association With Signal-Level Audiovisual Fusion. IEEE Transactions on Multimedia **6**(3) (2004) 406–413
37. Slaney, M., Covell, M.: FaceSync: A Linear Operator for Measuring Synchronization of Video Facial Images and Audio Tracks. Neural Information Processing Society **13** (2000)
38. Cutler, R., Davis, L.: Look Who's Talking: Speaker Detection using Video and Audio Correlation. International Conference on Multimedia and Expo (2000) 1589–1592
39. Nock, H., Iyengar, G., Neti, C.: Assessing Face and Speech Consistency for Monologue Detection in Video. Multimedia'02 (2002) 303–306
40. Iyengar, G., Nock, H., Neti, C.: Audio-Visual Synchrony for Detection of Monologues in Video Archives. International Conference on Acoustics, Speech, and Signal Processing (2003) 329–332
41. Chetty, G., Wagner, M.: "Liveness" Verification in Audio-Video Authentication. Australian International Conference on Speech Science and Technology (2004) 358–363
42. Sugamura, N., Itakura, F.: Speech Analysis and Synthesis Methods developed at ECL in NTT–From LPC to LSP. Speech Communications **5**(2) (1986) 199–215
43. Yehia, H., Rubin, P., Vatikiotis-Bateson, E.: Quantitative Association of Vocal-Tract and Facial Behavior. Speech Communication (28) (1998) 23–43
44. Bregler, C., Konig, Y.: "Eigenlips" for Robust Speech Recognition. International Conference on Acoustics, Speech, and Signal Processing **2** (1994) 19–22
45. Goecke, R., Millar, B.: Statistical Analysis of the Relationship between Audio and Video Speech Parameters for Australian English. International Conference on Audio-Visual Speech Processing (2003)
46. Eveno, N., Besacier, L.: Co-Inertia Analysis for "Liveness" Test in Audio-Visual Biometrics. International Symposium on Image and Signal Processing Analysis (2005) 257–261
47. Eveno, N., Besacier, L.: A Speaker Independent Liveness Test for Audio-Video Biometrics. 9th European Conference on Speech Communication and Technology (2005)
48. Chibelushi, C.C., Mason, J.S., Deravi, F.: Integrated Person Identification Using Voice and Facial Features. IEE Colloquium on Image Processing for Security Applications (4) (1997) 1–5
49. Smaragdis, P., Casey, M.: Audio/Visual Independent Components. International Symposium on Independent Component Analysis and Blind Signal Separation (2003) 709–714
50. Dolédec, S., Chessel, D.: Co-Inertia Analysis: an Alternative Method for Studying Species-Environment Relationships. Freshwater Biology **31** (1994) 277–294
51. Fisher, J.W., Darrell, T., Freeman, W.T., Viola, P.: Learning Joint Statistical Models for Audio-Visual Fusion and Segregation. Advances in Neural Information Processing Systems (2001)
52. Rabiner, L.R.: A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. In: IEEE. Volume 77. (1989) 257–286
53. Bengio, S.: An Asynchronous Hidden Markov Model for Audio-Visual Speech Recognition. Advances in Neural Information Processing Systems (2003)

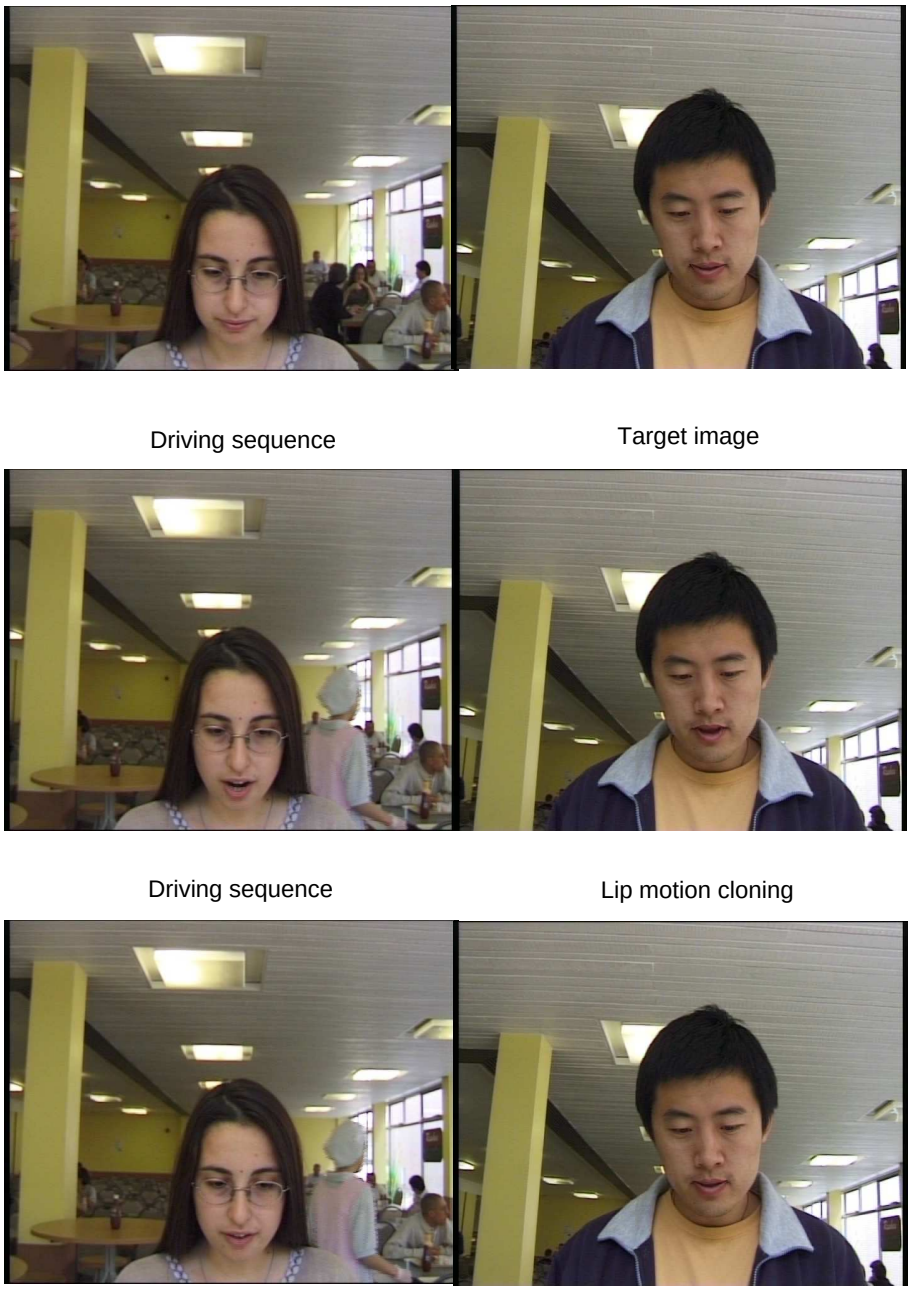

Fig. 4. Lip motion cloning from the driving sequence (left) to animate the static target image (right).