3

Audio-Visual Biometrics and Forgery

Hanna Greige and Walid Karam
University of Balamand
Lebanon

1. Introduction

With the emergence of smart phones and third and fourth generation mobile and communication devices, and the appearance of a "first generation" type of mobile PC/PDA/phones with biometric identity verification, there has been recently a greater attention to secure communication and to guaranteeing the robustness of embedded multi-modal biometric systems. The robustness of such systems promises the viability of newer technologies that involve e-voice signatures, e-contracts that have legal values, and secure and trusted data transfer regardless of the underlying communication protocol. Realizing such technologies require reliable and error-free biometric identity verification (IV) systems.

Biometric IV systems are starting to appear on the market in various commercial applications. However, these systems are still operating with a certain measurable error rate that prevents them from being used in a full automatic mode and still require human intervention and further authentication. This is primarily due to the variability of the biometric traits of humans over time because of growth, aging, injury, appearance, physical state, and so forth.

Imposture can be a real challenge to biometric IV systems. It is reasonable to assume that an impostor has knowledge of the biometric authentication system techniques used on one hand, and, on the other hand, has enough information about the target client (face image, video sequence, voice recording, fingerprint pattern, etc.) A deliberate impostor attempting to be authenticated by an IV system could claim someone else's identity to gain access to privileged resources. Taking advantage of the non-zero false acceptance rate of the IV system, an impostor could use sophisticated forgery techniques to imitate, as closely as possible, the biometric features of a genuine client.

The robustness of a biometric IV system is best evaluated by monitoring its behavior under impostor attacks. This chapter studies the effects of deliberate forgery on verification systems. It focuses on the two biometric modalities people use most to recognize naturally each other: face and voice.

The chapter is arranged as follows. **Section 2** provides a motivation of the research investigated in this chapter, and justifies the need for audio-visual biometrics. **Section 3** introduces audio-visual identity verification and imposture concepts. **Section 4** then reviews automated techniques of audio-visual (A/V) IV and forgery. Typically, an A/V IV system uses audio and video signals of a client and matches the features of these signals with stored templates of features of that client.

Next, **section 5** describes imposture techniques on the visual and audio levels: face animation and voice conversion. The video sequence of the client can be altered at the audio and the

visual levels. At the audio level, a voice transformation technique is employed to change the perceived speaker identity of the speech signal of the impostor to that of the target client. Techniques of voice transformation are surveyed. Such techniques were not originally developed for forgeries but have other applications. Voice conversion has been effectively used in text-to-speech systems to produce many different new voices. Other applications include dubbing movies and TV shows and the creation of virtual characters or even a virtual copy of a person's A/V identity. In this work, the main interest is to use voice conversion techniques in forgery.
At the visual level, a face transformation of the impostor to that of the client is required. This necessitates initially face detection and tracking, followed by face transformation. Face animation is described, which allows a 2-D face image of the client to be animated. This technique employs principal warps to deform defined MPEG-4 facial feature points based on determined facial animation parameters (FAP).
Evaluation and experimental results are then provided in **section 6**. Results of forgery are reported on the BANCA A/V database to test the effects of voice and face transformation on the IV system. The proposed A/V forgery is completely independent from the baseline A/V IV system, and can be used to attack any other A/V IV system. The Results drawn from the experiments show that state-of-the-art IV systems are vulnerable to forgery attacks, which indicate more impostor acceptance, and, for the same threshold, more genuine client denial. This should drive more research towards more robust IV systems, as outlined in the conclusion of **section 7**.

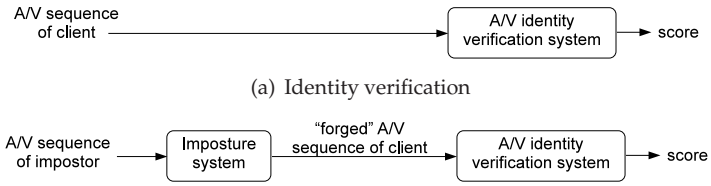
2. Why audio-visual biometrics?

The identification of a person is generally established by one of three schemes: Something the person owns and has (e.g. an identity card-ID, a security token, keys), something the person knows (e.g. username, password, personal identification number-PIN, or a combination of these), or something the person is or does, i.e. his/her anatomy, physiology, or behavior (e.g. fingerprint, signature, voice, gait). The latter is the more natural scheme and relies on biometric traits of a person for identification and authentication. An identity card can be lost, stolen or forged; a password or a PIN can be forgotten by its owner, guessed by an impostor, or shared by many individuals. However, biometric traits are more difficult to reproduce. Modern identity authentication systems have started to use biometric data to complement or even to replace traditional IV techniques.
The identity of a person is primarily determined visually by his or her face. A human individual can also fairly well be recognized by his or her voice. These two modalities, i.e. face and voice, are used naturally by people to recognize each other. They are also employed by many biometric identity recognition systems to automatically verify or identify humans for commercial, security and legal applications, including forensics. The combination of auditive and visual recognition is yet more efficient and improves the performance of the identity recognition systems.
Other biometric modalities have traditionally been used for identity recognition. Handwritten signatures have long been used to provide evidence of the identity of the signatory. Fingerprints have been used for over a hundred years to identify persons. They have also been successfully used in forensic science to identify suspects, criminals, and victims at a crime site. Other biometric features that help in identity recognition include the iris, the retina, hand geometry, vein pattern of hand, gait, electrocardiogram, ear form, DNA profiling, odor, keystroke dynamics, and mouse gestures. All of these biometric features, to the exception

of handwritten signatures, are considered intrusive; they require the cooperation of the user, careful exposure to biometric sensors, and might necessitate the repetition of feature capture for correctness and more accuracy. Some of these features, e.g. odor, keystroke dynamics, are not stable and vary from time to time, and thus are not reliable as an IV technique. DNA, on the other hand, is infallible but cannot be used instantly as it requires laboratory analysis.
Consequently, the adoption of the two non-intrusive, psychologically neutral modalities, i.e. face and voice, for automatic identity recognition is a natural choice, and is expected to mitigate rejection problems that often restrain the social use of biometrics in various applications, and broaden the use of A/V IV technologies.
In recent years, there has been a growing interest and research in A/V IV systems. Several approaches and techniques have been developed. These techniques are surveyed below. The robustness of these schemes to various quality conditions has also been investigated in the literature. However, the effect of deliberate imposture on the performance of IV systems has not been extensively reported. The robustness of a biometric IV system is best evaluated by monitoring its behavior under impostor attacks. Such attacks may include the transformation of one, many, or all of the biometric modalities.

3. Audio-visual identity verification and imposture

Typically, an automatic A/V IV system uses audio and video signals of a client and matches the features of these signals with stored templates of features of that client. A decision, in terms of a likelihood score, is made on whether to accept the claimed identity or to reject it. Fig. 1(a) depicts the concept. To be authenticated on a biometric system, an impostor attempts to impersonate a genuine client. He uses an imposture system to convert his own audio and video signals to defeat the verification system. This process is modeled in Fig. 1(b)



(a) Identity verification

(b) Imposture

Fig. 1. Audio-visual identity verification and imposture

The imposture concept can be summarized as follows. The video sequence of the client can be altered at the audio and the visual levels. At the audio level, a voice transformation technique is employed to change the perceived speaker identity of the speech signal of the impostor to that of the target client. At the visual level, a face transformation of the impostor to that of the client is required. This necessitates initially face detection and tracking, followed by face transformation. Fig. 2 depicts the concept.
The purpose of the work described in this chapter is to build such an A/V transformation imposture system and to provide a better understanding of its effect on the performance of automatic A/V IV systems. An attempt is made to increase the acceptance rate of the impostor and to analyzing the robustness of the recognition system.

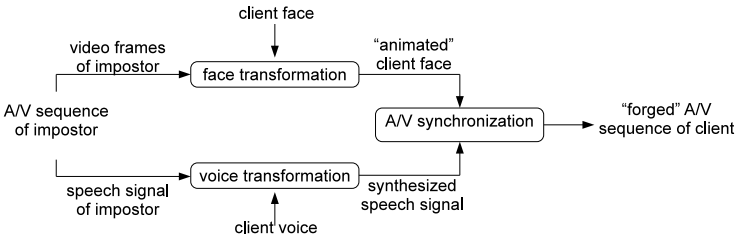


Fig. 2. The audio-visual imposture system

4. Audio-visual identity verification

An A/V IV system uses face and speech traits to verify (or to deny) a claimed identity. Face verification authenticates a person's identity by relying solely on facial information based on a set of face images (or a video sequence.) Speaker verification, on the other hand, authenticates the subject's identity based on samples of his speech. In this study, IV couples the speech and the face modalities by fusing scores of the respective verification systems.

4.1 Speaker verification

Speech carries primarily two types of information, the message conveyed by the speaker, and the identity of the speaker. In this work, analysis and synthesis of the voice of a speaker is text-independent and completely ignore the message conveyed. The focus is on the identity of the speaker.

To process a speech signal, a feature extraction module calculates relevant feature vectors from the speech waveform. On a signal window that is shifted at a regular rate a feature vector is calculated. Generally, cepstral-based feature vectors are used (section 4.1.1). A stochastic model is then applied to represent the feature vectors from a given speaker. To verify a claimed identity, 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 verification, the model should not reflect a specific speech structure, i.e. a specific sequence of words. State-of-the art systems use Gaussian Mixture Models (GMM) as stochastic models in text-independent mode (sections 4.1.2 and 4.1.3.)

Speaker verification encompasses typically two phases: a training phase and a test phase. During the training phase, the stochastic model of the speaker is calculated. The test phase determines if an unknown speaker is the person he claims to be. Fig. 3(a) and fig. 3(b) provide a block diagram representations of the concept.

4.1.1 Feature extraction

The first part of the speaker verification process is the speech signal analysis. Speech is inherently a non-stationary signal. Consequently, speech analysis is normally performed on short fragments of speech where the signal is presumed stationary. Typically, feature extraction is carried out on 20 to 30 ms windows with 10 to 15 ms shift between two successive windows. To compensate for the signal truncation, a weighting signal (Hamming window, Hanning windowÉ) is applied on each window of speech. The signal is also filtered with a first-order high-pass filter, called a pre-emphasis filter, to compensate for the -6dB/octave spectral slope of the speech signal. This pre-emphasis step is conventionally used before windowing.

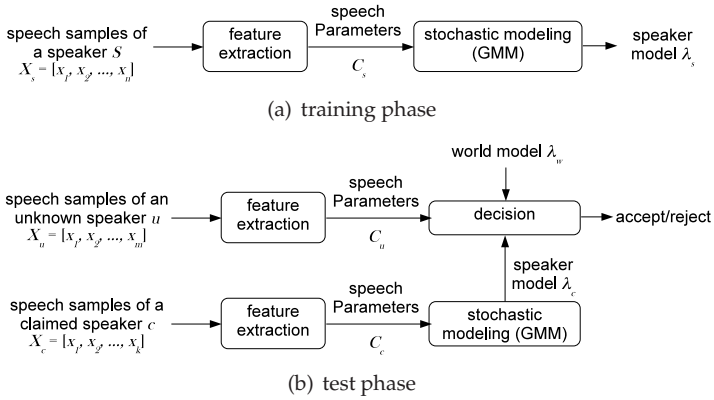


Fig. 3. The speaker verification system

Coding the truncated speech windows is achieved through variable resolution spectral analysis. Traditionally, two techniques have been employed: Filter-bank analysis, and linear-predictive analysis. Filter-bank analysis is a conventional spectral analysis technique that represents the signal spectrum with the log-energies using a filter-bank of overlapping band-pass filters. Linear predictive analysis is another accepted speech coding technique. It uses an all-pole filter whose coefficients are estimated recursively by minimizing the mean square prediction error.

The next step is cepstral analysis. The cepstrum is the inverse Fourier transform of the logarithm of the Fourier transform of the signal. A determined number of mel frequency cepstral coefficients (MFCC) are used to represent the spectral envelope of the speech signal. They are derived from either the filter bank energies or from the linear prediction coefficients. To reduce the effects of signals recorded in different conditions, Cepstral mean subtraction and feature variance normalization is used. First and second order derivatives of extracted features are appended to the feature vectors to account for the dynamic nature of speech.

4.1.2 Silence detection and removal

The silence part of the signal alters largely the performance of a speaker verification system. Actually, silence does not carry any useful information about the speaker, and its presence introduces a bias in the score calculated, which deteriorates the performance of the system. Therefore, most of the speaker recognition systems remove the silence parts from the signal before starting the recognition process. Several techniques have been used successfully for silence removal. In this work, we suppose that the energy in the signal is a random process that follows a bi-Gaussian model, a first Gaussian modeling the energy of the silence part and the other modeling the energy of the speech part. Given an utterance and more specifically the computed energy coefficients, the bi-Gaussian model parameters are estimated using the EM algorithm. Then, the signal is divided into speech parts and silence parts based on a maximum likelihood criterion. Treatment of silence detection is found in (Paoletti & Erten, 2000).

4.1.3 Speaker classification and modeling

Each speaker possesses a unique vocal signature that provides him with a distinct identity. The purpose of speaker classification is to exploit such distinctions in order to verify the

identity of a speaker. Such classification is accomplished by modeling speakers using a Gaussian Mixture Model (GMM).
Assume a given sample of speech Y, and a speaker S. Speaker verification is an attempt to determine if Y was spoken by S. The hypothesis test of equation 1 can be stated. An optimum test to decide between the null hypothesis H_0 and the alternative hypothesis H_1 is the likelihood ratio test:

$$\frac{p(Y|H_0)}{p(Y|H_1)} \begin{cases} \geq \theta & \text{accept } H_0 \\ < \theta & \text{reject } H_0 \end{cases} \tag{1}$$

$$\begin{cases} H_0\text{: Speech sample } Y \text{ belongs to speaker } S \\ H_1\text{: Speech sample } Y \text{ does not belong to speaker } S \end{cases}$$

where $p\left(Y \mid H_i\right),\ i = 0,1$ is the likelihood of the hypothesis H_i given the speech sample Y, and θ is the decision threshold for accepting or rejecting H_0. Figure 8 below describes speaker verification based on a likelihood ratio. A speech signal is first treated (noise reduction and linear filtering), then speaker-dependent features are extracted as described in section 4.1.1 above. The MFCC feature vectors, denoted $X = \{x_1, x_2, ..., x_n\}$, are then used to find the likelihoods of H_0 and H_1. Since speaker modeling is based on a mixture of Gaussians, as described in the next section 4.1.4, H_0 can be represented by the GMM model of the speaker λ_s, which symbolizes the mean vector and the covariance matrix parameters of the Gaussian distribution of the feature vectors of the speaker.

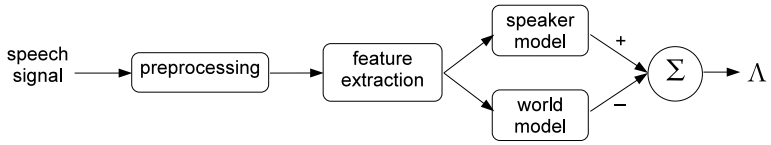


Fig. 4. Speaker verification based on a likelihood ratio

The alternative hypothesis H_1 can also be represented by a GMM model $\lambda_{\overline{s}}$, which models the entire space of alternatives to the speaker. This model is typically known as the "Universal Background Model" (UBM), or the "World Model". Alternative approaches represent the non-speaker space by a set of models representing the impostors.
The logarithm of the likelihood ratio $p\left(X \mid \lambda_s\right) / p\left(X \mid \lambda_{\overline{s}}\right)$ is often computed:

$$\Lambda(X) = \log p\left(X \mid \lambda_s\right) - \log p\left(X \mid \lambda_{\overline{s}}\right) \tag{2}$$

4.1.4 Gaussian mixture models

A mixture of Gaussians is a weighted sum of M Gaussian densities $P(x \mid \lambda) = \sum_{i=1:M} \alpha_i f_i(x)$ where x is an MFCC vector, $f_i(x)$ is a Gaussian density function, and α_i the corresponding weights. Each Gaussian is characterized by its mean μ_i and a covariance matrix Σ_i. A speaker model λ is characterized by the set of parameters $(\alpha_i, \mu_i, \sum_i)_{i=1:M}$.
For each client, two GMM's are used, the first corresponds to the distribution of the training set of speech feature vectors of that client, and the second represents the distribution of the training vectors of a defined "world model". To formulate the classification concept, assume

a speaker is presented along with an identity claim C. The feature vectors $X = \{\underline{x_i}\}_{i=1}^{N}$ are extracted. The average log likelihood of the speaker having identity C is calculated as

$$\mathcal{L}(X \mid \lambda_c) = \frac{1}{N}\sum_{i=1}^{N} \log p\left(\underline{x_i} \mid \lambda_c\right) \tag{3}$$

where

$$p\left(\underline{x_i} \mid \lambda_c\right) = \sum_{j=1}^{N_G} m_j \mathcal{N}\left(\underline{x}; \underline{\mu_j}, Cov_j\right)$$

$$\lambda_c = \left\{m_j \underline{\mu_j}, Cov_j\right\}_{j=1}^{N_G} \tag{4}$$

$$\mathcal{N}\left(\underline{x}; \underline{\mu_j}, Cov_j\right) = \frac{1}{(2\pi)^{\frac{D}{2}}\left|Cov_j\right|^{\frac{1}{2}}} e^{-\frac{1}{2}(\underline{x}-\underline{\mu_j})^T Cov_j^{-1}(\underline{x}-\underline{\mu_j})}$$

is a multivariate Gaussian function with mean $\underline{\mu_j}$ and diagonal covariance matrix Cov_j, and D is the dimension of the feature space, λ_c is the parameter set for person C, N_G is the number of Gaussians, m_j = weight for Gaussian j, and $\sum_{k=1}^{N_j} m_k = 1, m_k \geqslant 0 \forall k$. With a world model of w persons, the average log likelihood of a speaker being an impostor is found as $\mathcal{L}(X \mid \lambda_w) = \frac{1}{N}\sum_{i=1}^{N_W} \log p\left(\underline{x_i} \mid \lambda_w\right)$. An opinion on the claim is then found: $\mathcal{O}(X) = \mathcal{L}(X \mid \lambda_c) - \mathcal{L}(X \mid \lambda_w)$ As a final decision to whether the face belongs to the claimed identity, and given a certain threshold t, the claim is accepted when $\mathcal{O}(X) \geqslant t$, and rejected when $\mathcal{O}(X) < t$. To estimate the GMM parameters λ of each speaker, the world model is adapted using a Maximum a Posteriori (MAP) adaptation (Gauvain & Lee, 1994). The world model parameters are estimated using the Expectation Maximization (EM) algorithm (Dempster et al., 1977).

The EM algorithm is an iterative algorithm. Each iteration is 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 and therefore, the initialization of the model parameters is a crucial step.

GMM client training and testing is performed on the speaker verification toolkit BECARS (Blouet et al., 2004). BECARS implements GMM's with several adaptation techniques, e.g. Bayesian adaptation, MAP, maximum likelihood linear regression (MLLR), and the unified adaptation technique defined in (Mokbel, 2001).

The speaker recognition system requires two models: the model of the claimed speaker and the world model. The direct estimation of the GMM parameters using the EM algorithm requires a large amount of speech feature vectors. This can be easily satisfied for the world model where several minutes from several speakers may be collected offline. For the speaker model, this introduces a constraint, i.e. the speaker to talk for large duration. To overcome this, speaker adaptation techniques may be used (Mokbel, 2001) to refine the world model parameters λ_w into speaker specific parameters λ_s.

4.1.5 Score normalization

The last step in a speaker verification system is the decision of whether a speaker is the claimed identity (fig. 3(b)). It involves matching the claimed speaker model to speech samples by comparing a likelihood measure to a decision threshold. If the likelihood measure is smaller that the threshold, the speaker is rejected, otherwise, he is accepted. The choice of the decision threshold is not a simple task, and cannot be universally fixed due to the large score variability of various experiments. This is primarily due to the variability of conditions of speech capture, voice quality, speech sample duration, and background noise. It is also due to inter-speaker or intra-speaker differences between the enrollment speech data and the data used for testing. Score normalization is used to lessen the effect of the score variability problem. Different score normalization techniques have been proposed (Bimbot et al., 2004; Li & Porter, 1988). These include Z-norm, H-norm, T-norm, HT-norm, C-norm, D-norm, and WMAP.

Z-norm is a widely used normalization technique that has the advantage of being performed offline during the speaker training phase. Given a speech sample data X, a speaker model λ, and the corresponding score $L_\lambda(X)$. The normalized score is given by $\widetilde{L}_\lambda(X) = \frac{L_\lambda(X)-\mu_\lambda}{\sigma_\lambda}$, where μ_λ and σ_λ are the score mean and standard deviation of the speaker λ.

4.2 Face verification

The human face is a prominent characteristic that best distinguishes individuals. It forms an important location for person identification and transmits momentous information in social interaction. Psychological processes involved in face identification are known to be very complex, to be present from birth, and to involve large and widely distributed areas in the human brain. The face of an individual is entirely unique. It is determined by the size, shape, and position of the eyes, nose, eyebrows, ears, hair, forehead, mouth, lips, teeth, cheeks, chin, and skin.

Face verification is a biometric person recognition technique used to verify (confirm or deny) a claimed identity based on a face image or a set of faces (or a video sequence). Methods of face verification have been developed and surveyed in the literature (Chellappa et al., 1995; Dugelay et al., 2002; Fromherz et al., 1997; Zhang et al., 1997). (Zhao et al., 2003) classifies these methods into three categories: Holistic methods, feature-based methods, and hybrid methods. These methods are classified according to the differences in the feature extraction procedures and/or the classification techniques used.

4.2.1 Holistic methods

Holistic (global) identity recognition methods treat the image information without any localization of individual points. The face is dealt with as a whole without any explicit isolation of various parts of the face. Holistic techniques employ various statistical analysis, neural networks and transformations. They normally require a large training set but they generally perform better. However, such techniques are sensitive to variations in position, rotation, scale, and illumination and require preprocessing and normalization.

Holistic methods include Principal-component analysis (PCA) and eigenfaces (Turk & Pentland, 1991), Linear Discriminant Analysis (LDA) (Zhao et al., 1998), Support Vector Machine (SVM) (Phillips, 1998), and Independent Component Analysis (ICA) (Bartlett et al., 2002).

4.2.2 Feature-based methods

As opposed to holistic methods, feature-based techniques depend on the identification of fiducial points (reference or feature points) on the face such as the eyes, the nose, and the mouth. The relative locations of these feature points are used to find geometric association between them. Face recognition thus combines independent processing of the eyes, the nose, the mouth, and other feature points of the face. Since detection of feature points precedes the analysis, such a system is robust to position variations in the image.
Feature-based methods include graph matching (Wiskott et al., 1997), Hidden Markov Models (HMM) (Nefian & Hayes, 1998; Samaria & Young, 1994), and a Bayesian Framework (Liu & Wechsler, 1998; Moghaddam et al., 2000).

4.2.3 Hybrid and other methods

These methods either combine holistic and feature-based techniques or employ methods that do not fall in either category. These include Active Appearance Models (AAM) (Edwards et al., 1998), 3-D Morphable Models (Blanz & Vetter, 2003), 3-D Face Recognition (Bronstein et al., 2004), the trace transform (Kadyrov & Petrou, 2001; Srisuk et al., 2003), and kernel methods (Yang, 2002; Zhou et al., 2004).
The process of automatic face recognition can be thought of as being comprised of four stages: 1-Face detection, localization and segmentation, 2-Normalization, 3-Facial feature extraction, and 4-Classification (identification and/or verification). These subtasks have been independently researched and surveyed in the literature, and are briefed next.

4.2.4 Face detection and tracking in a video sequence

4.2.4.1 Face detection

Face detection is an essential part of any face recognition technique. Given an image, face detection algorithms try to answer the following questions

- Is there a face in the image?
- If there is a face in the image, where is it located?
- What are the size and the orientation of the face?

Face detection techniques are surveyed in (Hjelmas & Low, 2001). The face detection algorithm used in this work has been introduced initially by (Viola & Jones, 2001) and later developed further by (Lienhart & Maydt, 2002) at Intel Labs. It is a machine learning approach based on a boosted cascade of simple and rotated haar-like features for visual object detection.

4.2.4.2 Face tracking in a video sequence

Face tracking in a video sequence is a direct extension of still image face detection techniques. However, the coherent use of both spatial and temporal information of faces makes the detection techniques more unique. The technique used in this work employs the algorithm developed by (Lienhart & Maydt, 2002) on every frame in the video sequence. However, three types of errors are identified in a talking face video: 1-More than one face is detected, but only one actually exists in a frame, 2-A wrong object is detected as a face, and 3-No faces are detected. Fig. 5 shows an example detection from the BANCA database (Popovici et al., 2003), where two faces have been detected, one for the actual talking-face subject, and a false alarm. The correction of these errors is done through the exploitation of spatial and temporal

information in the video sequence as the face detection algorithm is run on every subsequent frame. The correction algorithm is summarized as follows:

(a) More than one face area detected: The intersections of these areas with the area of the face of the previous frame are calculated. The area that corresponds to the largest calculated intersection is assigned as the face of the current frame. If the video frame in question is the first one in the video sequence, then the decision to select the proper face for that frame is delayed until a single face is detected at a later frame, and verified with a series of subsequent face detections.

(b) No faces detected: The face area of the previous frame is assigned as the face of the current frame. If the video frame in question is the first one in the video sequence, then the decision is delayed as explained in part (a) above.

(c) A wrong object detected as a face: The intersection area with the previous frame face area is calculated. If this intersection ratio to the area of the previous face is less than a certain threshold, e.g. 80%, the previous face is assigned as the face of the current frame.

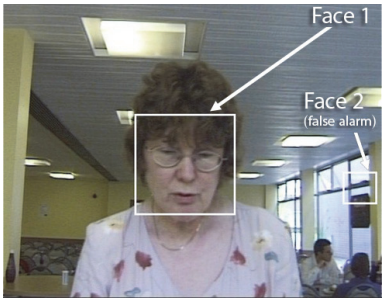


Fig. 5. Face Detection and Tracking

Fig. 6 below shows a sample log of face detection. The coordinates shown are (x, y) of the top left-hand corner of the rectangle encompassing the face area, and its width and height.

```
...
Frame 409: Found 1 face: 334 297 136 136 100.00 %
Frame 410: Found 1 face: 334 293 138 138 95.69 %
Frame 411: Found 1 face: 332 292 139 139 97.85 %
Frame 412: Found 2 faces. Previous face: 332 292 139 139
Face 0: 114 425 55 55 0.00 %
Face 1: 331 291 141 141 97.18 %
Selected face 1: 331 291 141 141

Frame 413: Found 1 face: 332 292 141 141 98.59 %
Frame 414: Found 1 face: 333 292 138 138 100.00 %
Frame 415: Found 1 face: 333 292 138 138 100.00 %
Frame 416: Found 1 face: 333 294 138 138 98.55 %
Frame 417: Found 3 faces. Previous face: 333 294 138 138
Face 0: 618 381 52 52 0.00 %
Face 1: 113 424 55 55 0.00 %
Face 2: 334 294 135 135 100.00 %
Selected face 2: 334 294 135 135
Frame 418: Found 1 face: 336 295 132 132 100.00 %
Frame 419: Found 1 face: 332 291 141 141 87.64 %
Frame 420: Found 1 face: 332 292 139 139 100.00 %
Frame 421: Found 1 face: 334 292 136 136 100.00 %
Frame 422: Found 1 face: 332 291 141 141 93.03 %
...
```

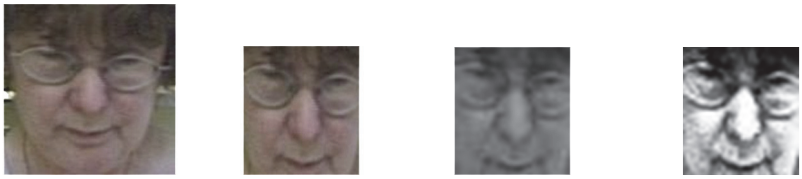
Fig. 6. Sample log of a face detection and tracking process

4.2.5 Face normalization

Normalizing face images is a required pre-processing step that aims at reducing the variability of different aspects in the face image such as contrast and illumination, scale, translation, rotation, and face masking. Many works in the literature (Belhumeur & Kriegman, 1998; Swets & Weng, 1996; Turk & Pentland, 1991) have normalized face images with respect to translation, scale, and in-plane rotation, while others have also included masking and affine warping to properly align the faces (Moghaddam & Pentland, 1997). (Craw & Cameron, 1992) have used manually annotated points around shapes to warp the images to the mean shape, leading to shape-free representation of images useful in PCA classification.
The pre-processing stage in this work includes four steps:

- Scaling the face image to a predetermined size (w_f, h_f).
- Cropping the face image to an inner-face, thus disregarding any background visual data.
- Disregarding color information by converting the face image to grayscale.
- Histogram equalization of the face image to compensate for illumination changes.

Fig. 7 below shows an example of the four steps.



(a) detected face (b) cropped "inner" face (c) face (d) histogram-equalized face

Fig. 7. Preprocessing face images

4.2.6 Facial feature extraction

The face feature extraction technique used in this work is DCT-*mod2*, initially introduced by (Sanderson & Paliwal, 2002), who showed that their proposed face feature extraction technique outperforms PCA and 2-D Gabor wavelets in terms of computational speed and robustness to illumination changes. This feature extraction technique is briefed next. A face image is divided into overlapping $N \times N$ blocks. Each block is decomposed in terms of orthogonal 2-D DCT basis functions, and is represented by an ordered vector of DCT coefficients:

$$\left[c_0^{(b,a)} c_1^{(b,a)} \ldots c_{M-1}^{(b,a)}\right]^T \tag{5}$$

where (b, a) represent the location of the block, and M is the number of the most significant retained coefficients. To minimize the effects of illumination changes, horizontal and vertical delta coefficients for blocks at (b, a) are defined as first-order orthogonal polynomial coefficients:

$$\Delta^h c_n^{(b,a)} = \frac{\sum_{k=-1}^{1} k h_k c_n^{(b,a+k)}}{\sum_{k=-1}^{1} h_k k^2} \qquad \Delta^v c_n^{(b,a)} = \frac{\sum_{k=-1}^{1} k h_k c_n^{(b+k,a)}}{\sum_{k=-1}^{1} h_k k^2} \tag{6}$$

The first three coefficients c_0, c_1, and c_2 are replaced in (5) by their corresponding deltas to form a feature vector of size M+3 for a block at (b, a):

$$\left[\Delta^h c_0 \Delta^v c_0 \Delta^h c_1 \Delta^v c_1 \Delta^h c_2 \Delta^v c_2 c_3 c_4 ... c_{M-1}\right]^T \tag{7}$$

In this study, neighboring blocks of 8 × 8 with an overlap of 50% is used. M, the number of retained coefficients is fixed at 15.

4.2.7 Classification

Face verification can be seen as a two-class classification problem. The first class is the case when a given face corresponds to the claimed identity (client), and the second is the case when a face belongs to an impostor. In a similar way to speaker verification, a GMM is used to model the distribution of face feature vectors for each person. The reader is referred back to sections 4.1.3 and 4.1.4 for a detailed treatment of identity classification and Gaussian mixture modeling.

4.3 Audio-visual data fusion

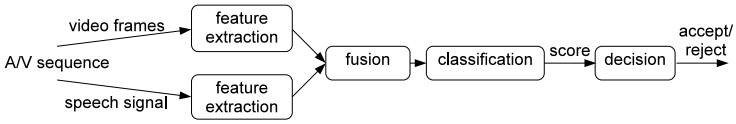
An A/V IV system uses face and voice biometric traits to verify (or to deny) a claimed identity. Face verification authenticates a person's identity by relying solely on facial information based on a set of face images (or a video sequence.) Speaker verification, on the other hand, authenticates the subject's identity based on samples of his speech. In this study, IV couples the speech and the face modalities by fusing scores of the respective verification systems.

It has been shown that biometric verification systems that combine multiple modalities outperform single biometric modality systems (Kittler, 1998). A final decision on the claimed identity of a person relies on both the speech-based and the face-based verification systems. To combine both modalities, a fusion scheme is needed. Fusion can be achieved at different levels (Ross & Jain, 2003) (Fig. 8):

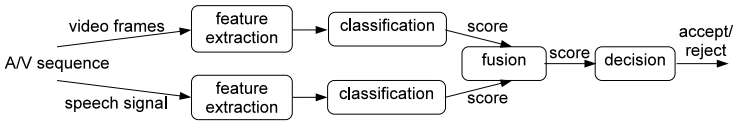
- Fusion at the feature extraction level, when extracted feature vectors originating from the multiple biometric systems are combined (Fig. 8(a))
- Fusion at the matching level, when the multiple scores of each system are combined (Fig. 8(b)), and
- Fusion at the decision level, when the accept/reject decisions are consolidated (Fig. 8(c)).

Different fusion techniques have been proposed and investigated in the literature. (Ben-Yacoub et al., 1999) evaluated different binary classification approaches for data fusion, namely Support Vector Machine (SVM), minimum cost Bayesian classifier, Fisher's linear discriminant analysis, C4.5 decision classifier, and multi layer perceptron (MLP) classifier. The use of these techniques is motivated by the fact that biometric verification is merely a binary classification problem. Other fusion techniques used include the weighted sum rule and the weighted product rule. It has been shown that the sum rule and support vector machines are superior when compared to other fusion schemes (Chatzis et al., 1999; Fierrez-Aguilar et al., 2003).

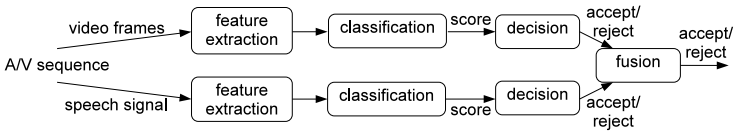
In this study, fusion at the classification level is used. The weighted sum rule fusion technique is employed to fuse the scores of the face and voice classifiers. The sum rule computes the A/V score s by weight averaging: $s = w_s s_s + w_f s_f$, where w_s and w_f are speech and face score weights computed so as to optimize the equal error rate (EER) on the training set. The



(a) Fusion at the feature extraction level



(b) Fusion at the classification level



(c) Fusion at the decision level

Fig. 8. The three levels of fusion of A/V biometric systems

speech and face scores must be in the same range (e.g. $\mu = 0, \sigma = 1$) for the fusion to be meaningful. This is achieved by normalizing the scores $s_{norm(s)} = \frac{s_s - \mu_s}{\sigma_s}$, $s_{norm(f)} = \frac{s_f - \mu_f}{\sigma_f}$.

5. Audio-visual imposture

Imposture is the act or conduct of a person (impostor) to pretend to be somebody else in an effort to gain financial or social benefits. In the context of this work, imposture is an attempt to increase the acceptance rate of one or more computer-based biometric verification systems and get authenticated to gain access to privileged resource. A/V imposture encompasses the transformation of both audio (voice) and visual (face) features, as developed below.

5.1 Voice transformation

Voice transformation, also referred to as speaker transformation, voice conversion, or speaker forgery, is the process of altering an utterance from a speaker (impostor) to make it sound as if it were articulated by a target speaker (client.) Such transformation can be effectively used by an avatar to impersonate a real human and converse with an Embodied Conversational Agent (ECA). Speaker transformation techniques might involve modifications of different aspects of the speech signal that carries the speaker's identity such as the Formant spectra i.e. the coarse spectral structure associated with the different phones in the speech signal (Kain & Macon, 1998), the Excitation function i.e. the "fine" spectral detail, the Prosodic features i.e. aspects of the speech that occur over timescales larger than individual phonemes, and the Mannerisms such as particular word choice or preferred phrases, or all kinds of other high-level behavioral characteristics. The formant structure and the vocal tract are represented by the overall spectral envelope shape of the signal, and thus are major features to be considered in voice transformation (Kain & Macon, 2001).

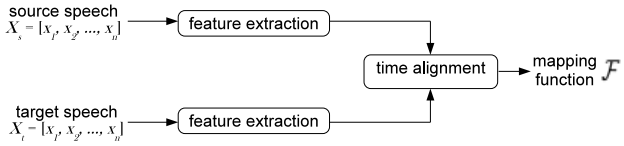
Several voice transformation techniques have been proposed in the literature (Abe et al. (1988); Kain & Macon (2001); Perrot et al. (2005); Stylianou & Cappe (1998); Sundermann et al. (2006); Toda (2009); Ye & Young (2004)). These techniques can be classified as text-dependent

methods and text independent methods. In text-dependent methods, training procedures are based on parallel corpora, i.e. training data have the source and the target speakers uttering the same text. Such methods include vector quantization (Abe et al. (1988); Arslan (1999)), linear transformation (Kain & Macon, 2001; Ye & Young, 2003), formant transformation (Turajlic et al., 2003), vocal tract length normalization (VTLN) (Sundermann et al., 2003), and prosodic transformation (Erro et al., 2010). In text-independent voice conversion techniques, the system trains on source and target speakers uttering different text. Techniques include text-independent VTLN (Sundermann et al., 2003), maximum likelihood adaptation and statistical techniques (Karam et al., 2009; Mouchtaris et al., 2004; Stylianou & Cappe, 1998), unit selection (Sundermann et al., 2006), and client memory indexation. (Chollet et al., 2007; Constantinescu et al., 1999; Perrot et al., 2005).

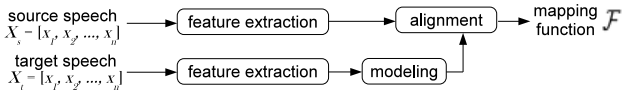
The analysis part of a voice conversion algorithm focuses on the extraction of the speaker's identity. Next, a transformation function is estimated. At last, a synthesis step is achieved to replace the source speaker characteristics by those of the target speaker.

Consider a sequence of spectral vectors uttered by a source speaker (impostor) $X_s = [x_1, x_2, ..., x_n]$, and another sequence pronounced by a target speaker $Y_t = [y_1, y_2, ..., y_n]$. Voice conversion is based on the estimation of a conversion function $\mathcal{F}$ that minimizes the mean square error $\epsilon_{mse} = E\lfloor \|y - \mathcal{F}(x)\|^2 \rfloor$, where E is the expectation. Two steps are useful to build a conversion system: training and conversion. In the training phase, speech samples from the source and the target speakers are analyzed to extract the main features. For text-dependent voice conversion, these features are time aligned and a conversion function is estimated to map the source and the target characteristics (Fig. 9(a)). In a text-independent voice conversion system, a model of the target speaker is used to estimate the mapping function as illustrated in Fig. 9(b).

The aim of the conversion is to apply the estimated transformation rule to an original speech pronounced by the source speaker. The new utterance sounds like the same speech pronounced by the target speaker, i.e. pronounced by replacing the target characteristics by those of the source voice. The last step is the re-synthesis of the signal to reconstruct the source speech voice (Fig. 10).



(a) in text-dependent voice conversion



(b) in text-independent voice conversion

Fig. 9. Voice conversion mapping function estimation

Voice conversion can be effectively used by an avatar to impersonate a real human and hide his identity in a conversation with an ECA. This technique is complementary with face transformation in the creation of an avatar that mimics in voice and face a target real human.

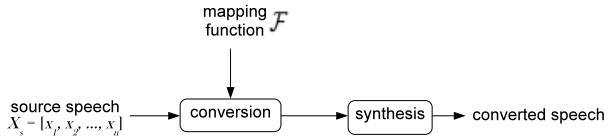


Fig. 10. Voice conversion

In this work, Mix*Trans*, initially introduced by (Karam et al., 2009), is used as a mixture-structured bias voice transformation, and it is briefed next.

5.1.1 Mix*Trans* speaker transformation

MixTrans is a text-independent speaker transformation technique that operates in the cepstral domain by defining a transformation that maps parts of the acoustical space to their corresponding time-domain signals. A speaker is stochastically represented with a GMM model. A deviation from this statistical estimate of the speaker it estimates could make the model better represent another speaker. Given a GMM model λ_t of a known target speaker t; λ_t can be obtained from a recorded speech of speaker t. The impostor s provides to the system the source speech X_s, which was never uttered by the target. Mix*Trans* makes use of X_s, its extracted features (MFCC), and the target model λ_t to compute the transformed speech X'_s. X'_s is time-aligned with, and has the same text as X_s, but appears to have been uttered by the target speaker t, as it inherits the characteristics of the source speaker s. Fig. 11 provides a block diagram of the proposed Mix*Trans* technique.

Mix*Trans* comprises several linear time-invariant filters, each of them operating in a part of the acoustical space:

$$\mathcal{T}_\gamma(\mathbf{X}) = \sum_k \mathbf{P}_k(\mathbf{X} + \mathbf{b}_k) = \sum_k \mathbf{P}_k\mathbf{X} + \sum_k \mathbf{P}_k\mathbf{b}_k = \mathbf{X} + \sum_k \mathbf{P}_k\mathbf{b}_k \tag{8}$$

where $\mathbf{b}_k$ represents the k^{th} bias and $\mathbf{P}_k$ is the probability of being in the k^{th} part of the acoustical space given the observation vector $\mathbf{X}$, parameter γ being the set of biases $\{\mathbf{b}_k\}$. $\mathbf{P}_k$ is calculated using a universal GMM modeling the acoustic space (Karam et al. (2009)). The parameters of the transformation are estimated such that source speech vectors are best represented by the target client GMM model λ using the maximum likelihood criterion $\hat{\gamma} = \arg\max_\gamma \mathcal{L}(\mathcal{T}_\gamma(\mathbf{X})|\lambda)$. Parameters $\{\mathbf{b}_k\}$ are calculated iteratively using the EM algorithm.

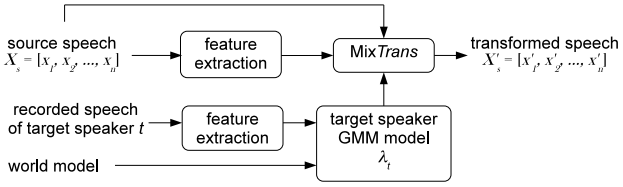


Fig. 11. Mix*Trans* block diagram

5.2 Face transformation

Face transformation is the process of converting someone's face to make it partially or completely different. Face transformation can be divided into two broad categories: Inter-person transformation, and intra-person transformation. In intra-person transformation,

the face is transformed in such a way as to have the subject retain his/her identity. The application of such transformation might be

- Face Beautification, i.e. to make the face look more attractive. Such applications are used in beauty clinics to convince clients of certain treatments. Example characteristic features of a female "more attractive" face include a narrower facial shape, fuller lips, a bigger distance of eyes, darker and narrower eye brows, a narrower nose, and no eye rings.
- Age modification, i.e. to make the face look younger or older by modifying the qualities of the facial appearance. Such qualities include those of a baby face, a juvenile, a teenager, and an adult. For example, the visual facial features of a baby include a small head, a curved forehead, large round eyes, small short nose, and chubby cheeks. One application of age modification is the projection of how a person might look when he gets old.
- Expression modification, i.e. the alteration of the mental state of the subject by changing the facial expression, e.g. joy, sadness, anger, fear, disgust, surprise, or neutral.
- Personalized Avatars, i.e. a computer user's representation of himself or herself, whether in the form of a 2-D or a 3-D model that could be used in virtual reality environments and applications such as games and online virtual worlds.

In inter-person transformation, the face is transformed so as to give the impression of being somebody else. The application of such transformation might be decoy, i.e. a form of protection for political, military, and celebrity figures. This involves an impersonator who is employed (or forced) to perform during large public appearances, to mislead the public. Such act would entail real time face transformation and 3-D animation and projection on large screens as well as imitating, naturally or synthetically, the public figure's voice, and mannerism.

- *Caricatural* impression, i.e. an exaggerated imitation or representation of salient features of another person in an artistic or theatrical way. This is used by impressionists, i.e. professional comedians, to imitate the behavior and actions of a celebrity, generally for entertainment, and makes fun of their recent scandals or known behavior patterns.
- Imposture or identity theft, i.e. the act of deceiving by means of an assumed character. The face of the impostor is altered so as to resemble some other existing person whose identity and facial features are publicly known in a certain real or virtual community. This application can be used to test the robustness of A/V IV systems to fraud. This latter concept is the purpose of this work.

Face transformation involves face animation of one or several pictures of a face, and can be two or three-dimensional. Face animation is described next.

5.2.1 Face animation

Computer-based face animation entails techniques and models for generating and animating images of the human face and head. The important effects of human facial expressions on verbal and non-verbal communication have caused considerable scientific, technological, and artistic interests in the subject. Applications of computer facial animation span a wide variety of areas including entertainment (animated feature films, computer games, etc.) communication (teleconferencing), education (distance learning), scientific simulation, and agent-based systems (e.g. online customer service representative).

To complete the scenario of A/V imposture, speaker transformation is coupled with face transformation. It is meant to produce synthetically an "animated" face of a target person, given a still photo of his face and some animation parameters defined by a source video sequence.
Several commercial and experimental tools for face animation are already available for the professional and the research communities. CrazyTalk[1] is an animation studio that provides the facility to create 3-D talking characters from photos, images or illustrations, and provides automatic lip-sync animation from audio and typed text. Other computer animated talking heads and conversational agents used in the research community include Greta (Pasquariello & Pelachaud, 2001), Baldi (Massaro, 2003), MikeTalk (Ezzat & Poggio, 1998), and Video Rewrite (Bregler et al., 1997), among others.
The face animation technique used in this work is MPEG-4 compliant, which uses a very simple thin-plane spline warping function defined by a set of reference points on the target image, driven by a set of corresponding points on the source image face. This technique is described next.

5.2.2 MPEG-4 2D face animation

MPEG-4 is an object-based multimedia compression standard, which defines a standard for face animation (Tekalp & Ostermann, 2000). It specifies 84 feature points 12 that are used as references for Facial Animation Parameters (FAPs). 68 FAPs allow the representation of facial expressions and actions such as head motion and mouth and eye movements. Two FAP groups are defined, visemes (FAP group 1) and expressions (FAP group 2). Visemes (FAP1) are visually associated with phonemes of speech; expressions (FAP2) are joy, sadness, anger, fear, disgust, and surprise.

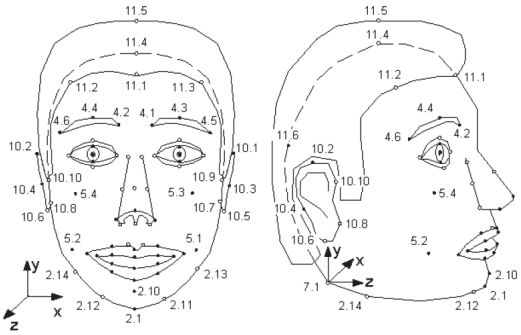


Fig. 12. MPEG-4 feature points

An MPEG-4 compliant system decodes a FAP stream and animates a face model that has all feature points properly determined. In this paper, the animation of the feature points is accomplished using a simple thin-plate spline warping technique, and is briefly described next.

5.2.3 Thin-plate spline warping

The thin-plate spline (TPS), initially introduced by Duchon (Duchon, 1976), is a geometric mathematical formulation that can be applied to the problem of 2-D coordinate

[1] http://www.reallusion.com/crazytalk/

transformation. The name thin-plate spline indicates a physical analogy to bending a thin sheet of metal in the vertical z direction, thus displacing x and y coordinates on the horizontal plane. TPS is briefed next. Given a set of data points $\{w_i, i = 1, 2, ..., K\}$ in a 2-D plane Ð for our case, MPEG-4 facial feature points Ð a radial basis function is defined as a spatial mapping that maps a location x in space to a new location $f(x) = \sum_{i=1}^{K} c_i \phi(\|x - w_i\|)$, where $\{c_i\}$ is a set of mapping coefficients and the kernel function $\phi(r) = r^2 \ln r$ is the thin-plate spline (Bookstein, 1989). The mapping function $f(x)$ is fit between corresponding sets of points $\{x_i\}$ and $\{y_i\}$ by minimizing the "bending energy" I, defined as the sum of squares of the second-order derivatives of the mapping function:

$$I[f(x,y)] = \iint_{\mathbb{R}^2} \left[\left(\frac{\partial^2 f}{\partial x^2} \right)^2 + 2 \left(\frac{\partial^2 f}{\partial xy} \right)^2 + \left(\frac{\partial^2 f}{\partial y^2} \right)^2 \right] dx\, dy \tag{9}$$

Fig. 13 shows thin-plate spline warping of a 2-D mesh, where point A on the original mesh is displaced by two mesh squares upward, and point B is moved downward by 4 mesh squares. The bent mesh is shown on the right of fig. 13.

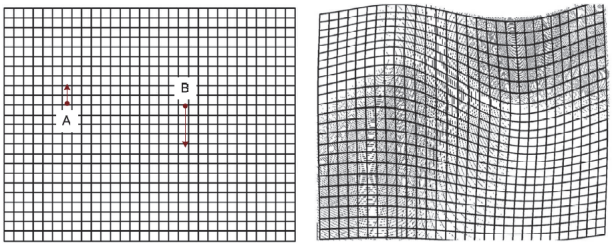


Fig. 13. TPS warping of a 2-D mesh

14 shows an example of TPS warping as it could be used to distort, and eventually animate a human face. The original face is on the left. The second and the third faces are warped by displacing the lowest feature point of the face, FDP number 2.1 as defined in the MPEG-4 standard 12. The corresponding FAP is "open_jaw". The last face on the right is a result of displacing the FDP's of the right eye so as to have the effect of a wink.

Fig. 14. Face animation using TPS warping

6. Evaluation and experimental results

To test the robustness of IV systems, a state-of-the-art baseline A/V IV system is built. This system follows the BANCA[2] "pooled test" of its evaluation protocol (Popovici et al., 2003).

[2] http://www.ee.surrey.ac.uk/CVSSP/banca/

The evaluation of a biometric system performance and its robustness to imposture is measured by the rate of errors it makes during the recognition process. Typically, a recognition system is a "comparator" that compares the biometric features of a user with a given biometric reference and gives a "score of likelihood". A decision is then taken based on that score and an adjustable defined acceptance "threshold" Θ. Two types of error rates are traditionally used: The False Acceptance Rate (FAR), i.e. the frequency that an impostor is accepted as a genuine client, and the False Rejection Rate (FRR), the frequency that a genuine client is rejected as an impostor. Results are reported in terms of an equal error rate (EER), the value where FAR=FRR. The lower the value of EER, the more performing the recognition system is. Typically, FAR and FRR can be traded off against each other by adjusting a decision threshold. The accuracy estimate of the EER is also reported by computing a confidence interval of 95% for the error rates.

6.1 Speaker verification

To process the speech signal, a feature extraction module calculates relevant feature vectors from the speech waveform. On a signal "FFT" window shifted at a regular rate, cepstral coefficients are derived from a filter bank analysis with triangular filters. A Hamming weighting window is used to compensate for the truncation of the signal. Then GMM speaker classification is performed with 128 and 256 Gaussians. The world model of BANCA is adapted using MAP adaptation, and its parameters estimated using the EM algorithm. A total of 234 true client tests and 312 "random impostor" tests per group were performed. EER's of 5.4%[$\pm$0.09] and 6.2%[$\pm$0.11] are achieved for 256 and 128 Gaussians respectively.

6.2 Face verification

Face verification is based on processing a video sequence in four stages: 1-Face detection, localization and segmentation, 2-Normalization, 3-Facial Feature extraction and tracking, and 4-Classification. The face detection algorithm used in this work is a machine learning approach based on a boosted cascade of simple and rotated haar-like features for visual object detection Lienhart & Maydt (2002). Once a face is detected, it is normalized (resized to 48x64, cropped to 36x40, gray-scaled, and histogram equalized) to reduce the variability of different aspects in the face image such as contrast and illumination, scale, translation, and rotation. The face tracking module extracts faces in all frames and retains only 5 per video for training and/or testing. The next step is face feature extraction. We use DCT-*mod*2 proposed in Sanderson & Paliwal (2002). In a similar way to speaker verification, GMM's are used to model the distribution of face feature vectors for each person. For the same BANCA "P" protocol, a total of 234 true clients and 312 "random impostor" tests are done (per group per frame, 5 frames per video.) EER's of 22.89%[$\pm$0.04] and 23.91%[$\pm$0.05] are achieved for 256 and 128 Gaussians respectively.

6.3 Speaker transformation

BANCA has total of 312 impostor attacks per group in which the speaker claims in his own words to be someone else. These attempts are replaced by the *MixTrans* transformed voices. For each attempt, MFCC analysis is performed and transformation coefficients are calculated in the cepstral domain using the EM algorithm. Then the signal transformation parameters are estimated using a gradient descent algorithm. The transformed voice signal is then reconstructed with an inverse FFT and OLA as described in Karam et al. (2009).

Verification experiments are repeated with the transformed voices. EER's of 7.88%[±0.08] and 7.96%[±0.10] are achieved for 256 and 128 Gaussians respectively.

6.4 Face transformation

Given a picture of the face of a target person, the facial feature points are first annotated as defined by MPEG-4. The facial animation parameters (FAP) used in the experiments correspond to a subset of 33 out of the 68 FAP's defines by MPEG-4. Facial actions related to head movement, tongue, nose, ears, and jaws are not used. The FAP's used correspond to mouth, eye, and eyebrow movements, e.g. horizontal displacement of right outer lip corner (stretch_r_corner_lip_o). A synthesized video sequence is generated by deforming a face from its neutral state according to determined FAP values, using the thin plate spline warping technique. Speaker verification experiments are repeated with the forged videos. EER's of 50.64%[±0.08] and 50.83%[±0.09] are achieved for 256 and 128 Gaussians respectively.

6.5 Audio-visual verification and imposture

Reporting IV results on A/V verification and A/V imposture is done simply by fusing scores of the verification of face and speaker and their transformations. In this paper, A/V scores are computed by weight averaging: $s = w_s s_s + w_f s_f$, where w_s and w_f are speech and face score weights computed so as to optimize EER on the training set, s_s and s_f being the speaker and the face scores respectively. Table 1 provides a summary of results of IV in terms of EER's, including speaker and face verification and their transformations, as well as A/V verification and transformation. The results clearly indicate an increase in EER's between the base experiments with no transformation and the experiments when either face, speaker, or both transformations are in effect. This indicates the acceptance of more impostors when any combination of voice/face transformation is employed.

		GMM size	
		256	128
speaker	no transformation	**5.40** [±0.09]	**6.22** [±0.11]
	MixTrans	7.88 [±0.08]	7.96 [±0.10]
face	no transformation	**22.89** [±0.04]	**23.91** [±0.05]
	TPS face warping	50.64 [±0.08]	50.83 [±0.09]
A/V	no transformation	**5.24** [±0.10]	**5.10** [±0.10]
	MixTrans+TPS warping	14.37 [±0.10]	15.39 [±0.10]
	MixTrans only	6.87 [±0.10]	6.60 [±0.10]
	TPS face warping only	13.38 [±0.10]	13.84 [±0.10]

Table 1. Summary of results - EER's with an accuracy estimate over a 95% interval of confidence

7. Conclusion

An important conclusion drawn from the experiments is that A/V IV systems are still far from being commercially feasible, especially for forensic or real time applications. A voice transformation technique that exploits the statistical approach of the speaker verification system can easily break that system and consent to a higher false acceptance rate. In a similar way to speaker verification systems, face verification systems that use the holistic

statistical approach are vulnerable to imposture attacks that exploit the statistical approach of the verification.

After decades of research and development on face and speech recognition, such systems are yet to find their place in our lives. With error rates of 1 to 5 percent for speech recognition, and 10 to 25 for face recognition, such systems have found their way only in exhibition proof-of-concept type of demos and limited noncritical applications, such as computer games and gadgets. It has remained an open question why A/V biometrics has remained in the research laboratory and has not found its way to public use. Will the world witness a breakthrough in A/V biometrics? Will we use our face and voice instead of our passport as we walk through the security zone at the airport to authenticate and declare our passage?

8. References

Abe, M., Nakamura, S., Shikano, K. & Kuwabara, H. (1988). Voice conversion through vector quantization, *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'08*, pp. 655–658.

Arslan, L. M. (1999). Speaker transformation algorithm using segmental codebooks (stasc), *Speech Commun.* 28(3): 211–226.

Bartlett, M., Movellan, J. & Sejnowski, T. (2002). Face recognition by independent component analysis, *Neural Networks, IEEE Transactions on* 13(6): 1450–1464.

Belhumeur, P. N. & Kriegman, D. J. (1998). What is the set of images of an object under all possible lighting conditions, *IJCV* 28: 270–277.

Ben-Yacoub, S., Abdeljaoued, Y. & Mayoraz, E. (1999). Fusion of Face and Speech Data for Person Identity Verification, *IEEE Transactions on Neural Networks* 10(05): 1065–1074.

Bimbot, F., Bonastre, J.-F., Fredouille, C., Gravier, G., Magrin-Chagnolleau, I., Meignier, S., Merlin, T., Ortega-Garcia, J., Petrovska-Delacretaz, D. & Reynolds, D. (2004). A tutorial on text-independent speaker verification, *EURASIP J. Appl. Signal Process.* 2004(1): 430–451.

Blanz, V. & Vetter, T. (2003). Face recognition based on fitting a 3d morphable model, *IEEE Trans. Pattern Anal. Mach. Intell.* 25(9): 1063–1074.

Blouet, R., Mokbel, C., Mokbel, H., Soto, E. S., Chollet, G. & Greige, H. (2004). Becars: A free software for speaker verification, *Proc. ODYSSEY'04*, pp. 145–148.

Bookstein, F. (1989). Principal warps: Thin-plate splines and the decomposition of deformations, *IEEE Transactions on Pattern Analysis and Machine Intelligence* 11(6): 567–585.

Bregler, C., Covell, M. & Slaney, M. (1997). Video rewrite: driving visual speech with audio, *SIGGRAPH '97: Proceedings of the 24th annual conference on Computer graphics and interactive techniques*, ACM Press/Addison-Wesley Publishing Co., New York, NY, USA, pp. 353–360.

Bronstein, A., Bronstein, M., Kimmel, R. & Spira, A. (2004). 3d face recognition without facial surface reconstruction, *ECCV'04: Proceesings of the European Conference on Computer Vision*, Prague.

Chatzis, V., Bors, A. G. & Pitas, I. (1999). Multimodal decision-level fusion for person authentication, *IEEE Transactions on Systems, Man, and Cybernetics, Part A* 29(6): 674–680.

Chellappa, R., Wilson, C. & Sirohey, S. (1995). Human and machine recognition of faces: a survey, *Proceedings of the IEEE* 83(5): 705–741.

Chollet, G., Hueber, T., Bredin, H., Mokbel, C., Perrot, P. & Zouari, L. (2007). *Advances in Nonlinear Speech Processing*, Vol. 1, Springer Berlin / Heidelberg, chapter Some experiments in Audio-Visual Speech Processing, pp. 28–56.

Constantinescu, A., Deligne, S., Bimbot, F., Chollet, G. & Cernocky, J. (1999). Towards alisp: a proposal for automatic language, *Independent Speech Processing. Computational Models of Speech Pattern Processing (ed. Ponting K.), NATO ASI Series*, pp. 375–388.

Craw, I. & Cameron, P. (1992). Face recognition by computer, *Proc. British Machine Vision Conference*, Springer Verlag, pp. 498–507.

Dempster, A. P., Laird, N. M. & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm, *Journal of the Royal Statistical Society* 39(1): 1–38.

Duchon, J. (1976). Interpolation des fonctions de deux variables suivant le principe de la flexion des plaques minces, *R.A.I.R.O. Analyse numérique* 10: 5–12.

Dugelay, J.-L., Junqua, J.-C., Kotropoulos, C., Kuhn, R., Perronnin, F. & Pitas, I. (2002). Recent advances in biometric person authentication, *ICASSP 2002 , 27th IEEE International Conference on Acoustics,Speech and Signal Processing - May 13-17, 2002, Orlando, USA*.

Edwards, G. J., Cootes, T. F. & Taylor, C. J. (1998). Face recognition using active appearance models, *ECCV '98: Proceedings of the 5th European Conference on Computer Vision-Volume II*, Springer-Verlag, London, UK, pp. 581–595.

Erro, D., Navas, E., Herná andez, I. & Saratxaga, I. (2010). Emotion conversion based on prosodic unit selection, *Audio, Speech, and Language Processing, IEEE Transactions on* 18(5): 974 –983.

Ezzat, T. & Poggio, T. (1998). Miketalk: A talking facial display based on morphing visemes, *Proceedings of the Computer Animation Conference*, pp. 96–102.

Fierrez-Aguilar, J., Ortega-Garcia, J., Garcia-Romero, D. & Gonzalez-Rodriguez, J. (2003). A comparative evaluation of fusion strategies for multimodal biometric verification, *Proceedings of IAPR International Conference on Audio and Video-based Person Authentication, AVBPA*, Springer, pp. 830–837.

Fromherz, T., Stucki, P. & Bichsel, M. (1997). A survey of face recognition, *Technical report*, MML Technical Report.

Gauvain, J.-L. & Lee, C.-H. (1994). Maximum a posteriori estimation for multivariate gaussian mixture observations of markov chains, *IEEE Transactions on Speech and Audio Processing* 2: 291–298.

Hjelmas, E. & Low, B. (2001). Face detection: A survey, *Computer Vision and Image Understanding* 83: 236–274(39).

Kadyrov, A. & Petrou, M. (2001). The trace transform and its applications, *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 23(8): 811–828.

Kain, A. & Macon, M. (1998). Spectral voice conversion for text-to-speech synthesis, *Acoustics, Speech and Signal Processing, 1998. Proceedings of the 1998 IEEE International Conference on* 1: 285–288 vol.1.

Kain, A. & Macon, M. (2001). Design and evaluation of a voice conversion algorithm based on spectral envelope mapping and residual prediction, *Acoustics, Speech, and Signal Processing, 2001. Proceedings. (ICASSP '01). 2001 IEEE International Conference on* 2: 813–816 vol.2.

Karam, W., Bredin, H., Greige, H., Chollet, G. & Mokbel, C. (2009). Talking-face identity verification, audiovisual forgery, and robustness issues, *EURASIP Journal on Advances in Signal Processing* 2009(746481): 18.

Kittler, J. (1998). Combining classifiers: a theoretical framework, *Pattern Analysis and Applications* 1: 18–27.

Li, K. P. & Porter, J. (1988). Normalizations and selection of speech segments for speaker recognition scoring, *In Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing*, pp. 595–598.

Lienhart, R. & Maydt, J. (2002). An extended set of haar-like features for rapid object detection, *2002 International Conference on Image Processing. 2002. Proceedings*, Vol. 1, pp. I–900–I–903.

Liu, C. & Wechsler, H. (1998). A unified bayesian framework for face recognition, *Image Processing, 1998. ICIP 98. Proceedings. 1998 International Conference on* 1: 151–155 vol.1.

Massaro, D. W. (2003). A computer-animated tutor for spoken and written language learning, *ICMI '03: Proceedings of the 5th international conference on Multimodal interfaces*, ACM, New York, NY, USA, pp. 172–175.

Moghaddam, B., Jebara, T. & Pentland, A. (2000). Bayesian face recognition, *Pattern Recognition 2000* 33(11): 1771–1782.

Moghaddam, B. & Pentland, A. (1997). Probabilistic visual learning for object representation, *IEEE Trans. Pattern Anal. Mach. Intell.* 19(7): 696–710.

Mokbel, C. (2001). Online adaptation of hmms to real-life conditions: A unified framework, *IEEE Transactions on Speech and Audio Processing* 9: 342–357.

Mouchtaris, A., der Spiegel, J. V. & Mueller, P. (2004). Non-parallel training for voice conversion by maximum likelihood constrained adaptation, *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'04*, Vol. 1, pp. I1–I4.

Nefian, A. & Hayes, M. I. (1998). Hidden markov models for face recognition, *Acoustics, Speech and Signal Processing, 1998. Proceedings of the 1998 IEEE International Conference on* 5: 2721–2724 vol.5.

Paoletti, D. & Erten, G. (2000). Enhanced silence detection in variable rate coding systems using voice extraction, *Circuits and Systems, 2000. Proceedings of the 43rd IEEE Midwest Symposium on* 2: 592–594 vol.2.

Pasquariello, S. & Pelachaud, C. (2001). Greta: A simple facial animation engine, *In Proc. of the 6th Online World Conference on Soft Computing in Industrial Applications*.

Perrot, P., Aversano, G., Blouet, R., Charbit, M. & Chollet, G. (2005). Voice forgery using alisp: Indexation in a client memory, *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'05*, Vol. 1, pp. 17–20.

Phillips, P. J. (1998). Support vector machines applied to face fecognition, *Adv. Neural Inform. Process. Syst* 1(11): 803–809.

Popovici, V., Thiran, J., Bailly-Bailliere, E., Bengio, S., Bimbot, F., Hamouz, M., Kittler, J., Mariethoz, J., Matas, J., Messer, K., Ruiz, B. & Poiree, F. (2003). The BANCA Database and Evaluation Protocol, *4th International Conference on Audio- and Video-Based Biometric Person Authentication, Guildford, UK*, Vol. 2688 of *Lecture Notes in Computer Science*, SPIE, Berlin, pp. 625–638.

Ross, A. & Jain, A. K. (2003). Information fusion in biometrics, *Pattern Recogn. Lett.* 24(13): 2115–2125.

Samaria, F. & Young, S. (1994). Hmm based architecture for face identification, *IntŠl J. Image and Vision Computing* 1(12): 537583.

Sanderson, C. & Paliwal, K. K. (2002). Fast feature extraction method for robust face verification, *IEE Electronics Letters* 38(25): 1648–1650.

Srisuk, S., Petrou, M., Kurutach, W. & Kadyrov, A. (2003). Face authentication using the trace transform, *Computer Vision and Pattern Recognition, 2003. Proceedings. 2003 IEEE Computer Society Conference on* 1: I–305–I–312 vol.1.

Stylianou, Y. & Cappe, O. (1998). A system for voice conversion based on probabilistic classification and a harmonic plus noise model, *Acoustics, Speech and Signal Processing, 1998. Proceedings of the 1998 IEEE International Conference on* 1: 281–284 vol.1.

Sundermann, D., Hoge, H., Bonafonte, A., Ney, H., Black, A. & Narayanan, S. (2006). Text-independent voice conversion based on unit selection, *Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on* 1: I–I.

Sundermann, D., Ney, H. & Hoge, H. (2003). Vtln-based cross-language voice conversion, *IEEE Automatic Speech Recognition and Understanding Workshop*, St. Thomas, Virgin Islands, USA, pp. 676–681.

Swets, D. L. & Weng, J. (1996). Using discriminant eigenfeatures for image retrieval, *IEEE Transactions on Pattern Analysis and Machine Intelligence* 18: 831–836.

Tekalp, A. & Ostermann, J. (2000). Face and 2-d mesh animation in mpeg-4, *Image Communication Journal* 15(4-5): 387–421.

Toda, T. (2009). Eigenvoice-based approach to voice conversion and voice quality control, *Proc. NCMMSC, International Symposium*, Lanzhou, China, pp. 492–497.

Turajlic, E., Rentzos, D., Vaseghi, S. & Ho, C. (2003). Evaluation of methods for parametric formant transformation in voice conversion, *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP'03)*.

Turk, M. & Pentland, A. (1991). Eigenfaces for recognition, *J. Cognitive Neuroscience* 3(1): 71–86.

Viola, P. & Jones, M. (2001). Rapid object detection using a boosted cascade of simple features, *IEEE CVPR'01*.

Wiskott, L., Fellous, J.-M., Kuiger, N. & von der Malsburg, C. (1997). Face recognition by elastic bunch graph matching, *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 19(7): 775–779.

Yang, M.-H. (2002). Kernel eigenfaces vs. kernel fisherfaces: Face recognition using kernel methods, *Automatic Face and Gesture Recognition, 2002. Proceedings. Fifth IEEE International Conference on* 1: 215–220.

Ye, H. & Young, S. (2003). Perceptually weighted linear transformations for voice conversion, *Proceedings of EUROSPEECH'03, Geneva*, Vol. 4, pp. 2409–2412.

Ye, H. & Young, S. (2004). Voice conversion for unknown speakers, *Proceedings of the ICSLPŠ04, Jeju Island, South Korea*, pp. 1161–1164.

Zhang, J., Yan, Y. & Lades, M. (1997). Face recognition: eigenface, elastic matching, and neural nets, *Proceedings of the IEEE*, pp. 1423–1435.

Zhao, W., Chellappa, R. & Krishnaswamy, A. (1998). Discriminant analysis of principal components for face recognition, *FG '98: Proceedings of the 3rd. International Conference on Face & Gesture Recognition*, IEEE Computer Society, Washington, DC, USA, p. 336.

Zhao, W., Chellappa, R., Phillips, P. J. & Rosenfeld, A. (2003). Face recognition: A literature survey, *ACM Comput. Surv.* 35(4): 399–458.

Zhou, S., Moghaddam, B. & Zhou, S. K. (2004). Intra-personal kernel space for face recognition, *In Proceedings of the IEEE International Automatic Face and Gesture Recognition*, pp. 235–240.