Pattern Anal Applic
DOI 10.1007/s10044-008-0121-2

THEORETICAL ADVANCES

Audio-visual speech asynchrony detection using co-inertia analysis and coupled hidden markov models

Enrique Argones Rúa · Hervé Bredin · Carmen García Mateo · Gérard Chollet · Daniel González Jiménez

Received: 8 February 2007 / Accepted: 2 April 2008

Abstract This paper addresses the subject of liveness detection, which is a test that ensures that biometric cues are acquired from a live person who is actually present at the time of capture. The liveness check is performed by measuring the degree of synchrony between the lips and the voice extracted from a video sequence. Three new methods for asynchrony detection based on co-inertia analysis (CoIA) and a fourth based on coupled hidden Markov models (CHMMs) are derived. Experimental comparisons are made with several methods previously used in the literature for asynchrony detection and speaker location. The reported results demonstrate the effectiveness and superiority of the proposed new methods based on both CoIA and CHMMs as asynchrony detection methods.

Keywords Biometrics · Video analysis · Statistics · Coupled hidden Markov models · Coinertia analysis · Video tracking

E. Argones Rúa (✉) · C. García Mateo · D. González Jiménez
SPG, STC Department, University of Vigo,
36200 Vigo, Spain
e-mail: eargones@gts.tsc.uvigo.es

C. García Mateo
e-mail: carmen@gts.tsc.uvigo.es

D. González Jiménez
e-mail: danisub@gts.tsc.uvigo.es

H. Bredin · G. Chollet
Dépt. TSI, CNRS-LTCI, GET-ENST, Paris, France
e-mail: bredin@tsi.enst.fr

G. Chollet
e-mail: chollet@tsi.enst.fr

1 Originality and contribution

This paper addressed the subject of liveness detection in frontal faces videos. The liveness check is performed by measuring the degree of synchrony between the lips and the voice extracted from a video sequence. Four different original methods are derived for that purpose: three methods based on co-inertia analysis and a fourth based on coupled hidden Markov models. The main contributions of this work are a full theoretical description of these methods and an experimental comparison of the main asynchrony detection algorithms in a publicly available database, allowing for future performance comparisons.

2 Introduction

Oral communication between people is a means of communication which is intrinsically multimodal. Not only does it include acoustic information but it also conveys complementary visual information. Acoustic information is classically used for state-of-the-art automatic speech processing applications such as automatic speech transcription or speaker authentication, while visual information is of great help in adverse environments where acoustic information is degraded (background noise, channel distortion, etc.). It provides complementary clues that can help in the analysis of the acoustic signal [1]. In extreme cases, visual information can even be used on its own. For instance, it is well known that deaf people can learn how to lip read. The joint analysis of acoustic and visual speech improves the robustness of automatic speech recognition systems [2, 3].

In the framework of identity verification based on talking-faces, most systems in the literature fuse scores from speaker verification and face recognition tests. Nevertheless, a

number of systems have attempted to make use of visual speech information to improve overall authentication performance [4–6].

One major weakness of these systems is that they do not take into account realistic impostor attack scenarios. Most existing systems, for example, could easily be fooled by simple attacks such as recording the voice of the target in advance and replaying it in front of the microphone, or simply placing a picture of the target's face in front of the camera. Another problem emerges in audio–visual speaker recognition when several faces appear in the video and the true speaker must be selected before identification or verification can take place. Systems such as the one described in [5] jointly model acoustic and visual speech in order to improve speaker verification performance with respect to independent modeling. The audio–visual biometric system described in [6] performs better when the visual stream is incorporated for both identification and verification. The robustness of these systems against non-synchronized video attacks or complex scenes with several face candidates, however, has not been tested.

One solution that has been proposed in the recent literature is to test liveness by studying the degree of synchrony between the acoustic signal and lip motion [7, 8]. Synchrony detection is not a new problem in audio–visual analysis. It is a major issue in fields such as speaker location [9] and speaker association [10–12]. Studies in the area used measures such as canonical correlation (CANCOR) [10] and mutual information (MI) [9, 11, 12] to distinguish the true speaker from a set of candidates. Synchrony detection in video-based biometrics would solve the problem of complex scenes where several faces are present in the image. Furthermore, it would allow the detection of attacks that cause audio–visual inconsistency. A number of studies in the biometrics field have already dealt with asynchrony detection. For instance, the method introduced in [7] fuses the speech and lip parameters in a single audiovisual feature vector stream, and then models it within a Gaussian mixture model (GMM) for each client. The results obtained with this method are impressive (1% equal error rate) for easy replay attacks constructed with a voice recording and a still photograph, although it has not been tested using a voice recording and an image sequence taken from another video. The method described in [8] uses co-inertia analysis (CoIA) correlation evolution to create liveness scores based on different delays between audio and image sequences.

The main aim of this paper is to describe a series of new asynchrony detection techniques and compare them to existing ones. The techniques presented increase the robustness of audio–visual biometric systems against spoof attacks. In addition to their application in the biometrics field, these techniques can also be applied to any generic audio–visual consistency assessment or monologue detection task. Two new approaches for measuring synchrony between audio and visual speech and detecting possible asynchrony are proposed. The first approach is based on co-inertia analysis (CoIA), and three new, different algorithms for detecting liveness are derived. The second one is a Bayesian approach based on coupled hidden Markov models (CHMMs). CANCOR, MI and the method proposed by Eveno et al. and based on CoIA are also tested in the same experimental framework for comparison purposes.

The rest of the paper is organized as follows. Section 2 introduces the acoustic and visual features that will be used in the experiments. The first approach (based on CoIA) is described in Sect. 3 and the second (based on CHMMs) in Sect. 4. A third method based on the fusion of the two previous approaches is investigated in Sect. 5. The methods used for comparison are introduced in Sect. 6. Finally, the performance of each of the methods is evaluated using real data from the BANCA audiovisual database. Evaluation protocols and results are discussed in Sect. 7.

3 Audiovisual speech features

3.1 Acoustic speech features

Mel-frequency cepstral coefficients (MFCC) are classical acoustic speech features in automatic speech processing. They are state-of-the-art features in many applications, including automatic speech recognition and speaker verification systems.

Every 10 ms, a 20 ms window is extracted from the acoustic signal and 12 MFCCs and the signal energy are computed to produce 13-dimensional acoustic speech features. First- and second-order time-derivatives are then appended, and finally a 39-dimensional feature vector is extracted every 10 ms.

3.2 Visual speech features

Visual speech features can be classified into two categories, depending on whether they are based on the shape or the appearance of the mouth [13]. The first category includes features that are directly related to the shape of the lip, such as the openness of the mouth, the location of particular lip landmarks, etc. The second category, in contrast, considers the mouth area as a whole and includes features that have been extracted directly from the pixels corresponding to a region of interest (ROI) around the mouth area.

Shape-based features Robust tracking of lip landmarks is a mandatory preliminary step towards extracting shape-based features. A *Lucas–Kanade*-based tracker [14] is used to track the location of a collection of facial landmarks

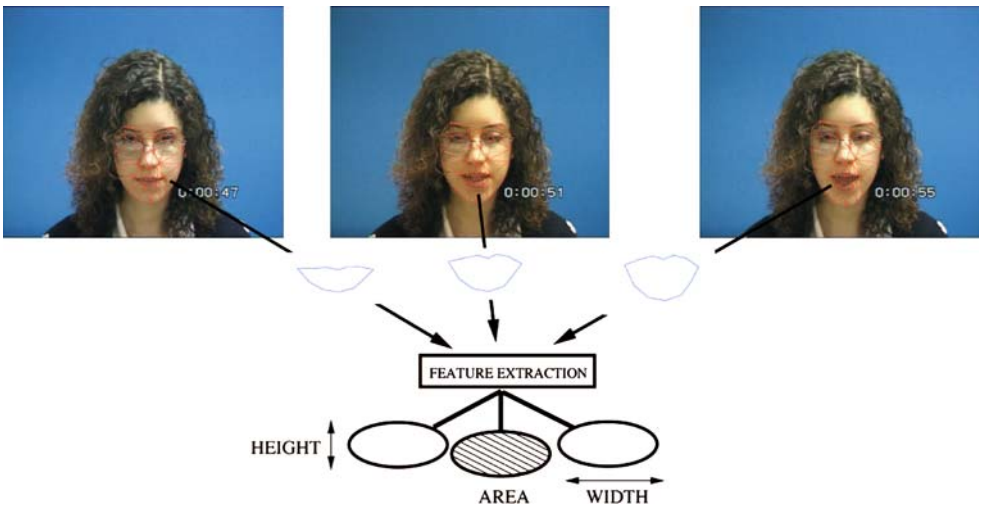


Fig. 1 Shape-based features extraction

(including lip landmarks) throughout the video sequence, as shown in the example in Fig. 1. Shape features corresponding to three separate dimensions (height, width and area of the mouth) are then straightforwardly extracted from the location of these lip landmarks.

Appearance-based features The mouth detection algorithm described in [15] was used to locate the lip area, as shown in Fig. 2. A discrete cosine transform (DCT) was then applied to the grey level size-normalized ROI, and the first 30 DCT coefficients (in a zig-zag manner, corresponding to the low spatial frequency) were kept as the visual speech features.

Sample rate The visual speech sample rate is dependent on the frame rate of the audiovisual sequence. Whereas current video cameras work at a frame rate of 25 or 29.97 frames/s (depending on the codec), the acoustic speech features presented in Sect. 2.1 are extracted at a sample rate of 100 Hz.

The algorithms presented here make use of acoustic and visual features that have equal sample rates. Therefore, the chosen solution was to linearly interpolate the visual features to obtain a sample rate of 100 Hz for both acoustic and visual features.

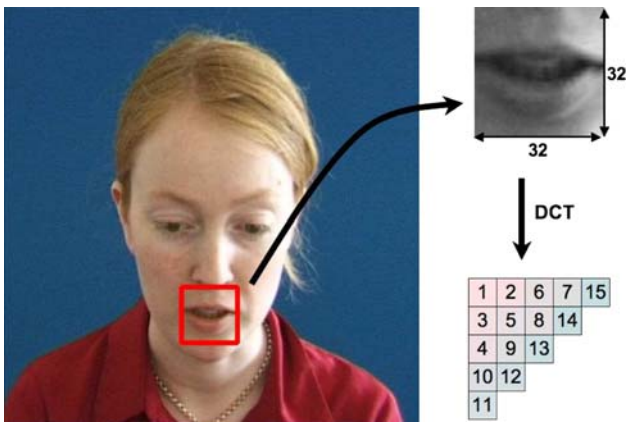


Fig. 2 Appearance-based features extraction

Visual dynamic features As with acoustic features, first- and second-order derivatives are also appended to static visual features. In the end, nine-dimensional shape-based features and 90-dimensional appearance-based features are available every 10 ms.

4 Coinertial approach: CoIA

4.1 Theoretical aspects

CoIA was first introduced by Dolédec and Chessel [16] in the field of biology to uncover the hidden relationships between species and their environment. Because, however, we did not find any demonstration of the co-inertia analysis in the literature, we have included the following demonstration:

Given two multivariate random variables $X = (X_1, \ldots, X_n)^t \in \mathbb{R}^n$ and $Y = (Y_1, \ldots, Y_m)^t \in \mathbb{R}^m$ of covariance matrix $C_{XY} = \mathrm{cov}(X, Y) = \mathrm{E}\{XY^t\} \in \mathbb{M}_{n\times m}$, where $\mathrm{E}\{\cdot\}$ denotes the expectation operator, CoIA allows to find $\mathbf{a} \in \mathbb{U}^n$ and $\mathbf{b} \in \mathbb{U}^m$, with $\mathbb{U}^l = \{z \in \mathbb{R}^l | \|z\| = 1\}$, so that the projections of X and Y on these two vectors have maximum covariance:

$$\begin{aligned}(\mathbf{a}, \mathbf{b}) &= \underset{(a,b)\in\mathbb{U}^n\times\mathbb{U}^m}{\mathrm{argmax}} \ \mathrm{cov}(a^tX, b^tY) \\ &= \underset{(a,b)\in\mathbb{U}^n\times\mathbb{U}^m}{\mathrm{argmax}} \ \mathrm{E}\{(a^tX)(Y^tb)\} \\ &= \underset{(a,b)\in\mathbb{U}^n\times\mathbb{U}^m}{\mathrm{argmax}} \ a^tC_{XY}b.\end{aligned} \tag{1}$$

Proposition 1 (CoIA) *$\boldsymbol{a}$ is the eigenvector corresponding to the highest eigenvalue λ of matrix $C_{XY}C_{XY}^t$ and $\boldsymbol{b}$ is proportional to $C_{XY}^t\,\boldsymbol{a}$.*

Proof of Proposition 1 Let us denote

$$\rho = a^t C_{XY} b \tag{2}$$

In the process of maximizing ρ, one can assume that $\rho > 0$ (change a into $-a$ if $\rho < 0$): it is therefore equivalent to maximize ρ and ρ^2.

$$\rho^2 = (a^t C_{XY} b)^t (a^t C_{XY} b) \tag{3}$$

$$\rho^2 = \left[\left(C_{XY}^t a\right)^t b \right]^t \left[\left(C_{XY}^t a\right)^t b \right]. \tag{4}$$

According to the Cauchy–Schwarz inequality, $\rho^2 \leq \|C_{XY}^t a\| \cdot \|b\|$ with equality if and only if b can be written as $\mu C_{XY}^t a$, with $\mu \in \mathbb{R}$. Therefore, Eq. 2 becomes:

$$\rho = a^t C_{XY} \left(\mu C_{XY}^t a\right) \tag{5}$$

$$\rho = \mu a^t \left(C_{XY} C_{XY}^t\right) a. \tag{6}$$

Since $\|a\| = 1, \rho$ is proportional to the Rayleigh quotient $R\left(C_{XY} C_{XY}^t, a\right) = (a^t C_{XY} C_{XY}^t a)/(a^t a)$, which is maximized when a is the eigenvector of $C_{XY}\, C_{XY}^t$ associated with the biggest eigenvalue λ_1.□

Sorting the eigenvalues of $C_{XY}\, C_{XY}^t$ in decreasing order $\{\lambda_1, \ldots, \lambda_d\}$, CoIA recursively finds the orthogonal vectors $\{\mathbf{a_1}, \ldots, \mathbf{a_d}\}$ and $\{\mathbf{b_1}, \ldots, \mathbf{b_d}\}$ which maximize the covariance between the projections $\mathbf{a_k^t} X$ and $\mathbf{b_k^t} Y$ (d being the rank of C_{XY}). In other words, CoIA rotates X and Y into a new coordinate system that maximizes their covariance.

In the following, $\mathbf{A}$ and $\mathbf{B}$ will denote $n \times d$ and $m \times d$ matrices containing the directions of the new coordinate systems:

$$\mathbf{A} = [\mathbf{a_1}|\ldots|\mathbf{a_d}] \quad \text{and} \quad \mathbf{B} = [\mathbf{b_1}|\ldots|\mathbf{b_d}].$$

4.2 Application of CoIA

4.2.1 *Extracting correlated acoustic and visual speech features*

Given *synchronized* acoustic and visual features $X \in \mathbb{R}^n$ and $Y \in \mathbb{R}^m$, CoIA can be used to compute matrices $\mathbf{A}$ and $\mathbf{B}$, which, in turn, can be used to extract *correlated* acoustic and visual features $\mathcal{X} = \mathbf{A}^t X$ and $\mathcal{Y} = \mathbf{B}^t Y$ of dimension d as follows:

$$\begin{aligned} \forall k \in \{1, \ldots, d\}, \quad \mathcal{X}_k &= \mathbf{a_k^t} X = \sum_{i=1}^{n} \mathbf{a_k}_i X_i \\ \mathcal{Y}_k &= \mathbf{b_k^t} Y = \sum_{i=1}^{m} \mathbf{b_k}_i Y_i. \end{aligned} \tag{7}$$

The effect of CoIA on real data is shown in Fig. 3, which contains features extracted from the audiovisual sequence"1002_f_g1_s02_1002_en.avi" from the BANCA database [17].

Remark CoIA can be used to reduce the dimension of acoustic and visual features without losing those that contain the most information regarding correlation. This is particularly important when working with CHMMs such as those described in Sect. 4. The curse of dimensionality is a major issue for these models because the small size of the BANCA database does not permit accurate training with high-dimensional features. The only requirement in our case was that all the acoustic and visual features X and Y had to be transformed using the same matrices $\mathbf{A}^{\Omega}$ and $\mathbf{B}^{\Omega}$.

Synchronized acoustic and visual features X^{Ω} and Y^{Ω} can be extracted from a training set Ω (BANCA world model part wm, in our case). CoIA transformation matrices $\mathbf{A}^{\Omega} = \left[\mathbf{a_1^{\Omega}}|\ldots|\mathbf{a_d^{\Omega}}\right]$ and $\mathbf{B}^{\Omega} = \left[\mathbf{b_1^{\Omega}}|\ldots|\mathbf{b_d^{\Omega}}\right]$ are then obtained by applying CoIA to Ω, and the transformed acoustic and visual features $\mathcal{X}^{\Omega}$ and $\mathcal{Y}^{\Omega}$ are computed using Eq. 8

$$\begin{aligned} \forall k \in \{1, \ldots, d\}, \quad \mathcal{X}_k^{\Omega} &= \mathbf{a}_{\mathbf{k}^{\Omega}}^t X = \sum_{i=1}^{n} \mathbf{a_k}_i^{\Omega} X_i \\ \mathcal{Y}_k^{\Omega} &= \mathbf{b}_{\mathbf{k}^{\Omega}}^t Y = \sum_{i=1}^{m} \mathbf{b_k}_i^{\Omega} Y_i. \end{aligned} \tag{8}$$

The dimensions of the transformed acoustic and visual features $\mathcal{X}^{\Omega}$ and $\mathcal{Y}^{\Omega}$ can then be conveniently reduced by keeping only the D most informative ones with respect to correlation.

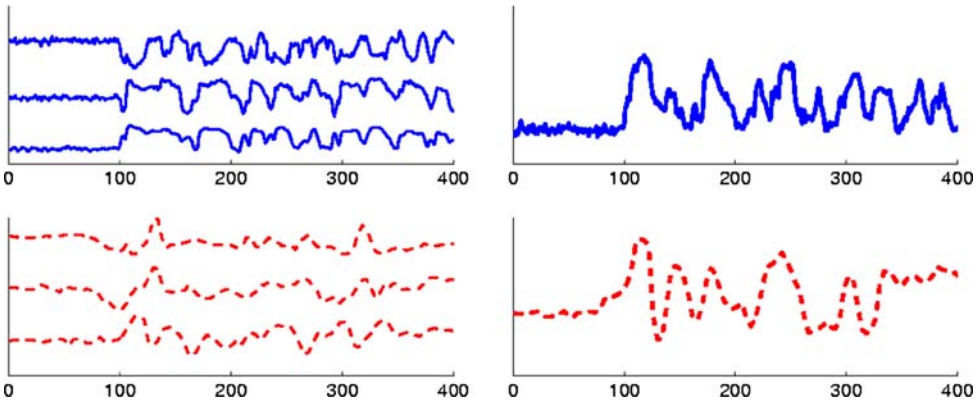


Fig. 3 Original acoustic and visual features (*top left*: X_1, X_2 and X_3 *bottom left* Y_1, Y_2 and Y_3) and first *correlated* acoustic and visual features (*top right*: $\mathcal{X}_1$ *bottom right* $\mathcal{Y}_1$). The correlation between X and Y is much more evident if we look at $\mathcal{X}_1$ and $\mathcal{Y}_1$

4.2.2 Measuring audiovisual speech synchrony

In this section, we introduce a method involving the use of *correlated* acoustic and visual features to measure how well voice X and lips Y correspond to each other. We distinguish between three different methods (world-, self- or piecewise self-training), though they all share a common framework:

1. The transformation matrices $\mathbf{A}^{\Omega}$ and $\mathbf{B}^{\Omega}$ are derived by means of CoIA from a training set Ω composed of acoustic and visual features X^{Ω} and Y^{Ω}.
2. Acoustic and visual features X^{Γ} and Y^{Γ} from a test utterance Γ are then transformed into $\mathcal{X}^{\Omega}$ and $\mathcal{Y}^{\Omega}$ using the previously computed matrixes $\mathbf{A}^{\Omega}$ and $\mathbf{B}^{\Omega}$:

$$\begin{aligned}\mathcal{X}^{\Omega} &= \mathbf{A}^{\Omega^t} X^{\Gamma} \\ \mathcal{Y}^{\Omega} &= \mathbf{B}^{\Omega^t} Y^{\Gamma}.\end{aligned} \tag{9}$$

3. Direct correlation is computed between each dimension of $\mathcal{X}^{\Omega}$ and $\mathcal{Y}^{\Omega}$ and used as a measure $s(X^{\Gamma}, Y^{\Gamma})$ of synchronization between X^{Γ} and Y^{Γ}, whereby the higher the correlation, the greater the synchronization:

$$\begin{aligned} s\left(X^{\Gamma}, Y^{\Gamma}\right) &= \frac{1}{D}\sum_{k=1}^{D} \frac{\mathcal{X}_k^{\Omega^t}\mathcal{Y}_k^{\Omega}}{\sqrt{\mathcal{X}_k^{\Omega^t}\mathcal{X}_k^{\Omega}}\sqrt{\mathcal{Y}_k^{\Omega^t}\mathcal{Y}_k^{\Omega}}} \\ &= \frac{1}{D}\sum_{\mathbf{k}=1}^{D} \frac{\left(\mathbf{a}_{\mathbf{k}}^{\Omega^t} X^{\Gamma}\right)^t \left(\mathbf{b}_{\mathbf{k}}^{\Omega^t} Y^{\Gamma}\right)}{\sqrt{\left(\mathbf{a}_{\mathbf{k}}^{\Omega^t} X^{\Gamma}\right)^t \left(\mathbf{a}_{\mathbf{k}}^{\Omega^t} X^{\Gamma}\right)}\sqrt{\left(\mathbf{b}_{\mathbf{k}}^{\Omega^t} Y^{\Gamma}\right)^t \left(\mathbf{b}_{\mathbf{k}}^{\Omega^t} Y^{\Gamma}\right)}}. \end{aligned} \tag{10}$$

The three methods mostly differ in how the training set Ω and the test set Γ are built.

World training method As proposed in the previous paragraph, one can use a large set of *synchronized* audiovisual sequences (the world model part wm of BANCA, in our case) to get X^{Ω} and Y^{Ω}. CoIA can then be used to compute matrices $\mathbf{A}^{\Omega}$ and $\mathbf{B}^{\Omega}$, modeling the *average best* correspondence between voice and lips. Using a given test utterance Γ, all the features in Γ are transformed using Eq. 9 to obtain $\mathcal{X}^{\Omega}$ and $\mathcal{Y}^{\Omega}$. A synchronization score $s(X^{\Gamma}, Y^{\Gamma})$ for test utterance Γ is then obtained using Eq. 10.

Self training method This method differs from the above in that a different training set is used to obtain matrices $\mathbf{A}^{\Omega}$ and $\mathbf{B}^{\Omega}$. Using a given test utterance Γ, CoIA is directly performed on data X^{Γ} and Y^{Γ}. In other words, the training and the test sets are the same: $\Gamma = \Omega$.

Piecewise self training method Bearing in mind that the purpose of this measure of synchronization is to discriminate between synchronized and non-synchronized audiovisual sequences, this third method is slightly different to the previous method. The intuition is the following (where a *sub-sequence* Λ is a sequence extracted from the original utterance sequence Γ by keeping only some of the samples, that is $\Lambda \subset \Gamma$) :

- if sequence Γ is synchronized, then every sub-sequence should follow the same synchronization model: a model $(\mathbf{A}^{\Omega}, \mathbf{B}^{\Omega})$ which is *optimal* with respect to a sub-sequence $\Omega \subset \Gamma$ would also be *optimal* with respect to any other sub-sequence $\Theta \subset \Gamma$;
- if the sequence is not synchronized, then a model $(\mathbf{A}^{\Omega},\mathbf{B}^{\Omega})$ which is *optimal* with respect to a sub-sequence Ω would not make sense for another sub-sequence $\Theta \subset \Gamma$ with $\Omega \cap \Theta = \varnothing$.

Let us introduce some notations:

- N is the number of samples in the sequence $\Gamma : X^{\Gamma} = \{x^1, \ldots, x^N\}$ and $Y^{\Gamma} = \{y^1, \ldots, y^N\}$.
- $\mathfrak{P}_{\Gamma}$ is the collection of all subsets of Γ of cardinal $\lfloor N/2 \rfloor$.

CoIA is applied to each training subsequence $\Omega \in \mathfrak{P}_{\Gamma}$ to produce transformation matrices $\mathbf{A}^{\Omega}$ and $\mathbf{B}^{\Omega}$. The remaining features in the sequence ($\Theta = \Gamma - \Omega$) are then transformed using the transformation matrices: $\mathcal{X}^{\Omega} = \mathbf{A}^{\Omega^t} X^{\Theta}$ and $\mathcal{Y}^{\Omega} = \mathbf{B}^{\Omega^t} Y^{\Theta}$. The synchronization measure $s(X^{\Theta}, Y^{\Theta})$ is computed as in Eq. 10, for every subsequence $\Omega \in \mathfrak{P}_{\Gamma}$. The final synchronization measure for sequence Γ is obtained via Eq. 11:

$$s(X^{\Gamma}, Y^{\Gamma}) = \frac{1}{\mathrm{card}\,\mathfrak{P}_{\Gamma}} \sum_{\Omega \in \mathfrak{P}_{\Gamma}} s\left(X^{\Theta}, Y^{\Theta}\right). \tag{11}$$

In practice, because it is not computationally feasible to use every $\Omega \in \mathfrak{P}$, only a few are drawn randomly (50, in our case) to compute the final synchronization measure.

5 Dynamic approach: CHMMs

5.1 Theoretical aspects

A CHMM can be seen as a collection of HMM in which the state at time t for every HMM in the collection is conditioned by the states of all the HMM in the collection at time $t-1$. This is illustrated in Fig. 4. The fact that the next state of every HMM depends on the states of all the HMMs is useful for capturing interactions between the acoustic and visual streams.

A CHMM can be completely described by the parameters $\boldsymbol{\lambda} = \{\lambda^i\} = \left\{\{\pi_j^i\}, \{a_{\mathbf{r}}^{i,s_i}\}, \{b_{s_i}^i(\cdot)\}\right\}$, for every stream $i \in \{1,\ldots,N_h\}$, where N_h is the number of streams; $s_i \in \{1,\ldots,NS_i\}$, where NS_i is the number of states in stream i; $\pi_{s_i}^i$ is the initial probability of the state s_i for stream i; $a_{\mathbf{r}}^{i,s_i}$ is the state transition probability for stream i and state s_i of

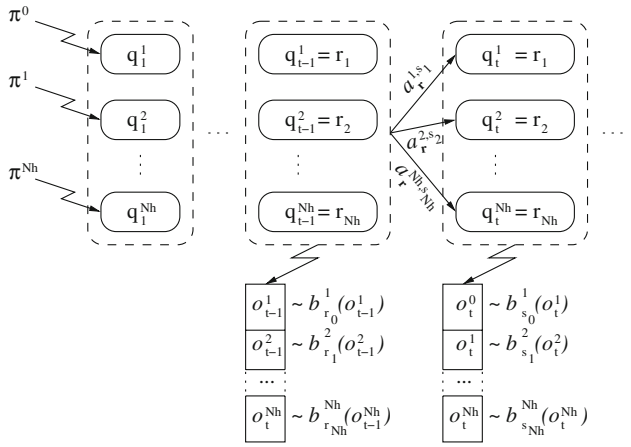


Fig. 4 The CHMM next state of each HMM depends on the state of all the HMM in the CHMM

the composite state $\mathbf{r} = \{r_1, \ldots, r_{N_h}\}$; and $b^i{}_{si}$ is the output probability density function for stream i and state s_i. The transition probabilities for stream i are defined as:

$$a^{i,s_i}_{\mathbf{r}} = \mathrm{P}(q^i_t = s_i | q^1_{t-1} = r_1, \ldots, q^{N_h}_{t-1} = r_{N_h}). \quad (12)$$

The output probability density function for every state s_i and stream i is modelled as gaussian mixture model (GMM) with $M^i_{s_i}$ mixtures. Let $o^i{}_t$ be the observation of the stream i at time t (in this case, $\mathbf{o^1} = \mathcal{X}^\Omega$ and $\mathbf{o^2} = \mathcal{Y}^\Omega$). The output probability density function can be written as:

$$b^i_{s_i}(o^i_t) = \mathrm{p}(o^i_t | q^i_t = s_i) = \sum_{m=1}^{M^i_{s_i}} w^i_{s_i,m} \mathcal{N}(o^i_t; \mu^i_{s_i,m}, \sigma^i_{s_i,m}). \quad (13)$$

The initial states for the training sequences are obtained using the five internal states of an energy-based voice activity detector (VAD) applied to the most *correlated* acoustic and visual features $\mathcal{X}_1$ and $\mathcal{Y}_1$, as defined in Eq. 7. Figure 6 shows the architecture of the VAD state machine, and Fig. 5 shows the VAD internal state sequence for a given signal. The VAD was chosen because it was believed that the system would be able to distinguish between synchronized and non-synchronized streams paying attention only to major signal changes (when a word starts or ends, when the signal is at a high energy interval, etc.) The state transition probabilities $a^{i,s_i}_{\mathbf{r}}$ are initially estimated from the state transitions obtained from the VAD sequence for all the training sequences:

$$a^{i,s_i}_{\mathbf{r}} = \frac{n^{i,s_i}_{\mathbf{r}}}{n^i_{\mathbf{r}}} \quad (14)$$

where $n^{i,s_i}_{\mathbf{r}}$ is the number of transitions to state s_i of stream i from the composite state $\mathbf{r} = \{r_1, \ldots, r_{N_h}\}$, and $n^i{}_{\mathbf{r}}$ is the total number of times that the CHMM visits the composite state $\mathbf{r} = \{r_1, \ldots, r_{N_h}\}$ before the last sample for every training sequence. The initial state probabilities $\pi^i_{s_i}$ can be estimated as $\pi^i_{s_i} = n^i_{s_i}/ns$, where $n^i_{s_i}$ are the number of training sequences in which the first state of stream i is state s_i, and ns is the total number of training sequences (Fig. 6).

It should be noted that the stream states are calculated independently for both streams in the training process. We can expect the output distribution for each stream and state to be the same (as if a HMM was trained on each stream separately) because no relation with the other stream's state is used for the initial state estimation. Dynamic relationships between the streams are then learnt from the combined state sequence of both streams.

The Baum–Welch algorithm adapted to the CHMM framework is iterated 20 times to train the CHMM. The Viterbi algorithm is used to calculate the sequence of states for every stream and the frame log-likelihoods. This framework has been derived in previous studies such as [2].

5.2 Bayesian framework to detect audiovisual asynchrony

In order to detect asynchrony between the acoustic and visual streams X and Y, a hypothesis test can be performed with the following hypothesis:

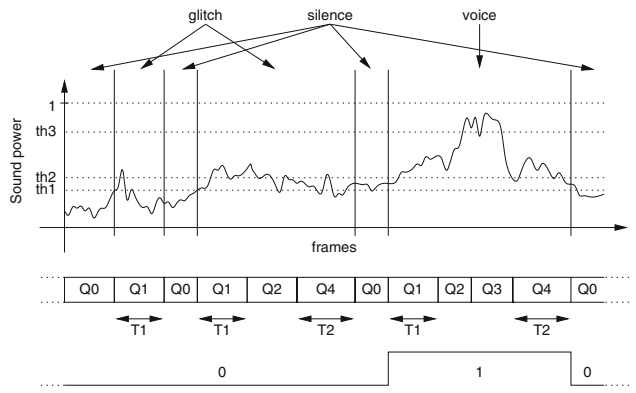


Fig. 5 Energy signal, VAD internal state sequence used to estimate the initial states for the training sequences and normal VAD output (voice/nonvoice)

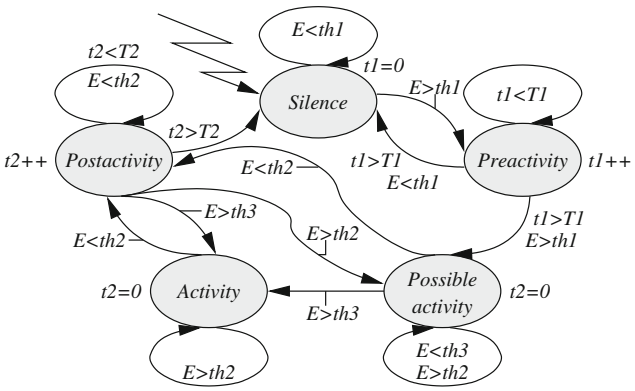


Fig. 6 VAD state machine. The only input variable is represented by E. Configurable timers $T1$ and $T2$ and thresholds $th1$, $th2$ and $th3$ can be tuned to modify the VAD behaviour

- $\mathcal{H}_0$: Because streams are produced synchronously the state sequences are dependent on each other. This hypothesis is represented by CHMM $\boldsymbol{\lambda}$.
- $\mathcal{H}_1$: Because streams are produced by independent sources, sequences are independent of each other. This hypothesis is represented by the two-stream HMM $\boldsymbol{\lambda}'$, as described in [3].

The test we performed in our study is a slight modification of the classical Bayesian test:

$$\mathcal{H}_0 \text{ is accepted} \iff \frac{\mathrm{p}(\mathcal{X},\mathcal{Y},\mathrm{Q}|\boldsymbol{\lambda})}{\mathrm{p}(\mathcal{X},\mathcal{Y},Q'|\boldsymbol{\lambda}')} > \theta, \tag{15}$$

where Q and Q' are the most likely state sequences. These likelihoods are provided by the Viterbi algorithm. This test approximates the classical Bayesian test when one state sequence is much more likely than the others. If the two-stream HMM $\boldsymbol{\lambda}' = \{\{\pi'^i_{s_i}\}, \{a'^{i,s_i}_{r_i}\}, \{b'^i_{s_i}(\cdot)\}\}$ was an independently trained model, then the slightest mismatch in the learned output distributions would thwart the effectiveness of the hypothesis test. In addition, dynamic relationships between the streams are encoded in the combined state sequences Q and Q'. The two-stream HMM $\boldsymbol{\lambda}'$ used in this hypothesis test, therefore, is an uncoupled version of the CHMM $\boldsymbol{\lambda}$, where the parameters for both the output distributions and the initial state probabilities are shared, and the state transition probabilities are computed from the CHMM $\boldsymbol{\lambda}$ parameters:

$$\left.\begin{array}{l} \pi'^i_{s_i} = \pi^i_{s_i} \\ b'^i_{s_i}(\cdot) = b^i_{s_i}(\cdot) \end{array}\right\} \quad \forall\, i \in \{1,\ldots,N_h\}, s_i \in \{1,\ldots,NS_i\}. \tag{16}$$

This enhances the asynchrony discrimination because random effects derived from the output probability density functions training are removed and only differences in the decoded state sequences are taken into account: if $\mathcal{H}_1$ holds then it is likely that rare joint state transitions in Q makes the ratio in Eq. 15 fall below θ. The state transition matrix of $\boldsymbol{\lambda}'$ is defined in such a way that the next state s_i for every HMM i depends only on its previous state r_i. It is known that:

$$\begin{aligned} a'^{i,s_i}_{r_i} &= \mathrm{P}(q^i_t = s_i | q^i_{t-1} = r_i) \\ &= \sum_{\mathbf{q_{t-1}}|q^i_{t-1}=r_i} \mathrm{P}(q^i_t = s_i|\mathbf{q_{t-1}} = \mathbf{r}) \prod_{j=1,j\neq i}^{N_h} \mathrm{P}(q^j_{t-1} = r_j) \\ &= \sum_{r_1=1}^{NS_1} \cdots \sum_{r_{i-1}=1}^{NS_{i-1}} \sum_{r_{i+1}=1}^{NS_{i+1}} \cdots \sum_{r_{N_h}=1}^{NS_{N_h}} a^{i,s_i}_{\mathbf{r}} \prod_{j=1,j\neq i}^{N_h} \mathrm{P}(q^j_{t-1} = r_j). \end{aligned} \tag{17}$$

The probability $\mathrm{P}(q^i{}_t = r_i)$ can be calculated. It depends on time, however, and it is not desirable to work with time-dependent state transition probabilities. Therefore, since the quantity $\lim_{t\to\inf} \mathrm{P}(q^i_t = r_i)$ converges quickly for ergodic HMMs, it is computed following this iterative procedure:

1. Initialization: for $t = 1$,

$$\mathrm{P}(q^i_1 = s_i) = \pi^i_{s_i}. \tag{18}$$

2. Induction:

$$\mathrm{P}(q^i_t{=}s_i) = \sum_r a^{i,s_i}_r \prod_{j=1}^{N_h} \mathrm{P}(q^j_{t-1}{=}r_j) \tag{19}$$

3. Stop condition:

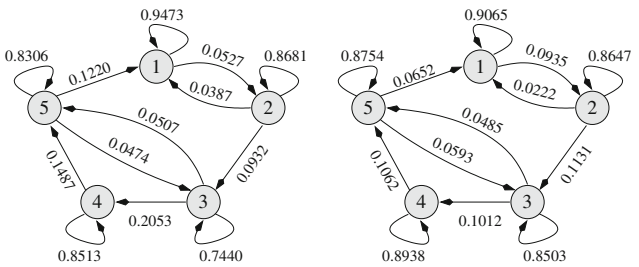
$$\left|\frac{\mathrm{P}(q^i_t = s_i) - \mathrm{P}(q^i_{t-1} = s_i)}{\mathrm{P}(q^i_t = s_i)}\right| < 10^{-6}. \tag{20}$$

An example of the uncoupled transition matrices obtained by this uncoupling procedure is illustrated in Fig. 7. It should be noted that the original CHMM from which the uncoupled transition matrices are obtained has 250 different $\{a^{i,s_i}_{\mathbf{r}}\}$ parameters.

6 Bayesian fusion using GMM as a probability density function estimator

CoIA and CHMM are different approaches to asynchrony detection. While CoIA uses linear correlation as a measure of synchrony between acoustic and visual features, CHMM uses dynamic statistics to determine whether acoustic and visual features are synchronous. Because they use complementary information, fusing them could lead to improved performance. Statistical fusion techniques such as GMM fusion [18] can be used for this purpose. In our framework, the joint probability density function f of the CoIA and CHMM scores s_1 and s_2 for both the synchronized $\mathfrak{S}$ and non-synchronized $\mathfrak{N}$ acoustic and visual features is modeled using two GMMs:

$$f_{\mathfrak{S}}(\mathbf{s}(X,Y)) = \mathrm{P}(\mathbf{s}(X,Y)|X \text{ and } Y \text{ are synchronized}) \tag{21}$$



Audio stream HMM after the uncoupling procedure

Video stream HMM after the uncoupling procedure

Fig. 7 Uncoupled HMMs obtained with the uncoupling procedure described in Sect. 4.2

$$f_{\mathfrak{N}}(\mathbf{s}(X,Y)) = \mathrm{P}(\mathbf{s}(X,Y)|X \text{ and } Y \text{are not synchronized}), \tag{22}$$

where $\mathbf{s}(X,Y) = (s_1(X,Y), s_2(X,Y))^t$ and $f_{\mathfrak{S}}$ and $f_{\mathfrak{N}}$ can both be expressed as follows:

$$f_\rho(\mathbf{s}) = \sum_{i=1}^{N} w_i^\rho \frac{1}{\sqrt{(2\pi)^d \|\Gamma_i^\rho\|}} \exp\left(-\frac{1}{2}(\mathbf{s}-\mu_i^\rho)^T \Gamma_i^{\rho-1}(\mathbf{s}-\mu_i^\rho)\right) \tag{23}$$

$f_{\mathfrak{S}}$ and $f_{\mathfrak{N}}$ are initialized using the LBG algorithm and trained using the EM algorithm. To discriminate between synchronized and not synchronized acoustic and visual streams, the following hypothesis test is performed:

$$X \text{ and } Y \text{ are synchronized} \iff \frac{f_{\mathfrak{S}}(\mathbf{s}(X,Y))}{f_{\mathfrak{N}}(\mathbf{s}(X,Y))} > \theta. \tag{24}$$

We used the above method in our fusion experiments described below. Additional results using the sum rule are reported as a baseline for fusion [19].

7 Other methods for asynchrony detection

Several asynchrony detection techniques have already been studied in the literature, as indicated in the introduction. We performed the same experiments with CHMM, CoIA approaches (including Eveno and Besacier's approach), CANCOR and MI to compare performance. Although descriptions of these approaches can be found in the literature [8–12], some implementation issues must be discussed in order to facilitate understanding of the results presented later in this paper.

Eveno's measure In a similar liveness test framework [8], *Eveno and Besacier* apply CANCOR analysis and CoIA to the tested sequence, in order to obtain the first projection vectors $\mathbf{a}_1$ and $\mathbf{b}_1$. The design of their synchrony measure $M(X,Y)$ (summarized by Eq. 27) results from the observation of the value of the correlation ρ between $\mathbf{a}_1^t X$ and $\mathbf{b}_1^t Y$, as a function of the shift δ between audio and visual features: its maximum value ρ_{ref} is often obtained for a small negative shift:

$$\rho_{\text{ref}} = \max_{-80\,\text{ms} \le \delta \le 0} \left[\text{corr}\left(\mathbf{a}_1{}^t X^\delta, \mathbf{b}_1{}^t Y\right)\right] \tag{25}$$

$$\rho_{\text{avg}} = \text{mean}\left[\text{corr}\left(\mathbf{a}_1{}^t X^\delta, \mathbf{b}_1{}^t Y\right)\right] \tag{26}$$

where X^δ is the δ-shifted X.

$$M(X,Y) = \frac{1}{2\Delta+1}\left(\frac{\rho_{\text{ref}}}{\rho_{\text{avg}}} - 1\right) \sum_{\delta=-\Delta}^{\Delta} f(\text{corr}(a_1^t X^\delta, b_1^t Y)) \tag{27}$$

where $f(\rho) = 1$ if $\rho \le \rho_{\text{ref}}$ and 0 otherwise, and Δ corresponds to a time-shift of 400 ms (10 visual frames). $M(X, Y)$ can be seen as a measure of the *peakiness* of the maximum found in the interval [−80, 0 ms]. Our implementation of *Eveno*'s algorithm involves the use of slightly different acoustic features to those described in [8] (MFCC instead of LPC). The major difference between our self-training method and Eveno's approach is that we considered more than just the first dimension, $\mathcal{X}_1$ and $\mathcal{Y}_1$. Moreover, the world-training method is also quite different in that it makes use of a prior training step where universal transformation matrices $\mathbf{A}^\Omega$ and $\mathbf{B}^\Omega$ are learned.

CANCOR CANCOR analysis is applied to synchrony detection [10] in the same manner as CoIA is. All the synchrony detection techniques described for CoIA can be directly tested using the CANCOR approach. The same training sets are used for the estimation of CoIA and CANCOR transformation matrices.

MI Mutual information between visual and acoustic parameters can be defined in several ways depending on the probability density estimator used to model joint and separate feature vectors. In our case we use GMMs as parameter estimators for visual, acoustic and joint visual and acoustic features in the CoIA-transformed space; the MI measure is therefore defined as:

$$MI(X,Y) = \sum_{t=1}^{N} f_{AV}(\mathcal{X}_t, \mathcal{Y}_t) \log\left(\frac{f_{AV}(\mathcal{X}_t, \mathcal{Y}_t)}{f_A(\mathcal{X}_t) f_V(\mathcal{Y}_t)}\right), \tag{28}$$

where f_{AV}, f_A and f_V are the GMM probability density functions, as defined in Eq. 23, for the joint audio–visual features, the audio features and the visual features, respectively.

8 Experiments

8.1 Experimental framework

BANCA database We conducted our experiments using the English part of the BANCA database [17], which was originally designed for biometric system evaluation purposes only. Two disjoint groups of 26 people (13 male and 13 female) recorded 24 sequences of approximately 15 s, in which they each pronounced a sequence of 10 digits and either their name and address (client access) or the name and address of another person (impostor access). The recordings were performed under three different conditions (controlled, degraded and adverse) as shown in Fig. 8. Sixty additional sequences from 30 different people were also recorded (under controlled, degraded and adverse conditions) and used to create the world model.

Fig. 8 Three different recording conditions. *Left* controlled (DV camera), *middle* degraded (webcam), *right* adverse (background noise)

Evaluation protocols Because we are focusing on asynchrony detection in this paper, the experimental protocols described in [17]—designed for identity verification—are not valid here. As a matter of fact, for each group, all the 312 (26 × 12) original client access sequences were synchronized naturally. Therefore, for each group, 3,432 (26 × 12 × 11) asynchronous recordings were built artificially using audio and video from two different recordings, in which the name and address pronounced were the same, both acoustically and visually. Two asynchrony detection protocols were derived from these two sets of synchronized and non-synchronized audiovisual sequences:

Controlled Only recordings from the controlled conditions are used. This protocol can be used to compare the suitability of both shape-based and appearance-based visual speech features for asynchrony detection. Only the controlled part of the world model recordings of BANCA can be used to train models. As a result, for each group, 104 synchronized and 312 non-synchronized sequences were tested using this protocol.

Pooled The 3 conditions (controlled, adverse and degraded) were used. This protocol can be used to estimate the robustness of CoIA and CHMM asynchrony detection methods. All the world model recordings of BANCA can be used to train models. As a result, for each group, 312 synchronized and 3,432 non-synchronized sequences were tested using this protocol.

Although it is very unlikely that an impostor would own both an audio and a video recording of the client pronouncing two different utterances, these protocols deal with an extremely challenging, if not the most challenging, synchrony detection task and therefore constitute a useful framework in which to compare the performance of the different synchrony measures we propose.

Performance measure and comparison Given a decision threshold θ, an asynchrony detection system can commit two types of error: it can falsely accept a non-synchronized sequence and classify it as a synchronized sequence (false acceptance) or it can falsely reject a synchronized sequence and classify it as a non-synchronized sequence (false rejection). A low θ value would tend to increase the number of false acceptances (FA) and reciprocally a high θ value would tend to increase the number of false rejections (FR). Consequently we defined the false acceptance rate (FAR) and false rejection rate (FRR) as a function of θ (one objective being to find the best compromise between those two error rates):

$$\mathrm{FAR}(\theta) = \frac{\mathrm{FA}(\theta)}{\mathrm{NI}} \text{ and } \mathrm{FRR}(\theta) = \frac{\mathrm{FR}(\theta)}{\mathrm{NC}} \tag{29}$$

where NI and NC are the numbers of non-synchronized and synchronized sequences respectively. Detection error tradeoff (DET) curves are usually plotted to compare such detection algorithms [20]. The (FAR(θ), FRR(θ)) point is plotted for every possible θ value and the resulting curve can be used to easily compare two systems: the closer the target curve is to the origin, the better.

Depending on the application, we might want to place more or less importance on false rejection or acceptance errors. The weighted error rate (WER), presented in [21], is therefore introduced:

$$\mathrm{WER}(r) = \frac{1}{1+r}(r \cdot \mathrm{FAR} + \mathrm{FRR}). \tag{30}$$

Two possible applications where mentioned in the introduction. Although the synchrony detection can be performed using the same algorithms in both applications, different compromises between FAR and FRR should be assumed, and hence we should choose different values for the weight r:

$r = 10$ This configuration corresponds to a biometric authentication system with strict security requirements, where the most important constraint is to detect spoof attacks. It is therefore ten times more costly to falsely accept a non-synchronized sequence than to reject a synchronized sequence (in that case, a genuine client would have to repeat his/her access attempt).

$r = 1$ This configuration might be used in an application where no strong binary decision (synchronized vs. non-synchronized) is needed. It could be used, for example, to select the true speaker form a large group of people on a screen.

Are results generalizable and conclusive? Because the BANCA database is divided into two disjoint groups, namely G1 and G2, the WERs for one group (the test set) are calculated using the thresholds that minimizes the WERs for the other group (the training set). This prevents the results from being biased by the choice of threshold. Confidence intervals at 95% are then computed using the method proposed in [22] with the following formula, where $\alpha = 1.960$ and $\overline{WER(r)}$ is the estimation obtained thanks to the test set:

$$\mathrm{WER}(r) \in \overline{\mathrm{WER}(r)} \pm \alpha \cdot \sqrt{\frac{1}{1+r^2}} \cdot \sqrt{\frac{r^2}{NI} \cdot \overline{\mathrm{FAR}}(1-\overline{\mathrm{FAR}}) + \frac{1}{NC} \cdot \overline{\mathrm{FRR}}(1-\overline{\mathrm{FRR}})}. \quad (31)$$

As a matter of fact, given the small number of tests that are performed, it is important to make sure that the resulting difference between the error rates of two methods is stastically significant and capable of generating conclusive results.

8.2 Experimental results

Table 1 shows the asynchrony detection performance of the different methods compared in this paper in terms of the WER (1.0) and WER (10). All the experiments were performed using both shape-based (shp) and appearance-based (app) visual features. Algorithms based on CANCOR, CHMM, MI and CoIA were used for the asynchrony detection. The *Method* column indicates the audiovisual synchrony measurement method used in the correlation-based CANCOR and CoIA cases. This column indicates the design parameters regarding stream dimension and number of gaussians in the case of MI and CHMM algorithms.

The DET curves for the different algorithms are shown in Figs. 9 and 10 (controlled and pooled protocols, respectively). It must be noted that parameters such as the stream dimension or the number of gaussians per state (in the case of the CHMM) may slightly alter the performance of these methods for the same dataset. These parameters have been empirically chosen to achieve a good compromise in terms of performance. In the case of the CHMM and MI approaches, the dimension of the streams used and the number of gaussians are shown in Table 1. CoIA and CANCOR methods use correlated acoustic and visual streams of dimension 3. In other words, $D = 3$ in Eq. 10 (Fig. 11).

Shape-based versus appearance-based visual features Performance was much better for appearance-based visual features than shape-based visual features in all of the methods (with no exceptions). This suggests that shape-based visual features do not hold appropriate linear dependencies with acoustic speech features or time evolution information. As a matter of fact, only the outer lip contours were modeled by the lip tracker: the area, height and width of the mouth and their time derivatives do not provide enough information for synchrony analysis. Appearance-based features, in contrast, contain (in a hidden way) not only the shape of the mouth but also additionnal information such as whether the mouth is really open, the tongue or the teeth are visible, etc.

CANCOR versus CoIA While WT performed far better than (P)ST for the CANCOR-based synchrony measure, CoIA-based measures did not behave in the same way and (P)ST methods yielded better results in all cases. This observation coincides with the findings of [8]. CANCOR needs much more training data to accurately estimate the transformation matrices **A** and **B**: world-training (where a lot of training data is available) therefore results in better modeling than self-training does (where only the sequence itself can be used for training). CoIA is much less dependent on the amount of training data available and it is even better at modeling and uncovering the intrinsical synchrony of a given audiovisual sequence.

CHMM robustness against degraded conditions A quick comparison of the performance of CoIA and CHMM using the controlled and pooled protocols shows that CHMM performs better than CoIA in degraded test conditions. The *WER* (1.0) of ST appearance-based CoIA increased from WER (1.0) = 8.25% for the controlled protocol to *WER* (1.0) = 11.9% for the pooled protocol (statistically significant degraded performance). Comparatively, we observed a small, yet not statistically significant, degradation in the performance of appearance-based CHMM. This observation was also made in the security-oriented performance measure defined by WER (10). This CHMM robustness against low quality features is highlighted when observing the CHMM performance when using the less informative shape-based features, where it gets the best performance.

Piecewise self-training One of the contributions of this paper is the introduction of the piecewise self-training approach. It seems to be particularly effective for applications where more security is needed [defined by the error rate WER (10)], and where conditions are controlled. Indeed, in such circumstances, piecewise self-training

Table 1 WER (1), WER (10) and their 95% confidence intervals (in subscripts) for the different algorithms for the controlled (C) and pooled (P) protocol using appearance (app) and shape-based (shp) visual parameters (VP)

Protocol	VP	Algorithm	Method	WER% (1.0)	WER% (10.0)
C	shp	CANCOR	WT	$22.76_{(17.94,\ 27.58)}$	$5.26_{(4.34,\ 6.18)}$
			ST	$25.72_{(20.69,\ 30.75)}$	$7.24_{(5.90,\ 8.58)}$
			PST	$28.12_{(23.01,\ 33.24)}$	$9.22_{(7.78,\ 10.66)}$
		CoIA	WT	$23.48_{(18.56,\ 28.39)}$	$5.67_{(4.87,\ 6.47)}$
			ST	$18.11_{(13.51,\ 22.70)}$	$5.13_{(4.21,\ 6.05)}$
			PST	$22.92_{(18.08,\ 27.75)}$	$6.66_{(5.47,\ 7.84)}$
			Eveno	$25.64_{(20.73,\ 30.55)}$	$9.35_{(8.52,\ 10.19)}$
		MI	*D*3, *ng*256	$41.91_{(36.50,\ 47.31)}$	$9.15_{(8.81,\ 9.49)}$
		CHMM	*D*3, *ng*4	$16.59_{(12.18,\ 20.99)}$	$6.06_{(4.54,\ 7.58}$
	app	CANCOR	WT	$10.18_{(6.53,\ 13.82)}$	$2.71_{(1.91,\ 3.51)}$
			ST	$13.14_{(9.24,\ 17.04)}$	$6.91_{(5.27,\ 8.54)}$
			PST	$20.35_{(15.59,\ 25.11)}$	$6.95_{(5.58,\ 8.32)}$
		CoIA	WT	$13.22_{(9.12,\ 17.32)}$	$3.12_{(2.25,\ 3.99)}$
			ST	$8.25_{(4.89,\ 11.61)}$	$2.74_{(1.72,\ 3.76)}$
			PST	$8.81_{(5.25,\ 12.38)}$	$2.32_{(1.55,\ 3.09)}$
			Eveno	$21.39_{(16.93,\ 25.85)}$	$9.43_{(8.47,\ 10.39)}$
		MI	*D*4, *ng*256	$40.71_{(35.23,\ 46.18)}$	$9.02_{(8.64,\ 9.39)}$
		CHMM	*D*4, *ng*8	$9.21_{(5.64,\ 12.79)}$	$2.71_{(1.77,\ 3.65)}$
		Fusion	Sum rule	$7.61_{(3.42,\ 9.24)}$	$2.01_{(1.14,\ 2.18)}$
			GMM	$6.33_{(3.42,\ 9.24)}$	$3.99_{(2.50,\ 5.48)}$
P	app	CANCOR	WT	$13.39_{(11.18,\ 15.61)}$	$3.69_{(3.28,\ 4.10)}$
			ST	$20.16_{(17.77,\ 22.56)}$	$7.82_{(7.36,\ 8.29)}$
			PST	$22.76_{(20.04,\ 25.48)}$	$6.75_{(6.27,\ 7.22)}$
		CoIA	WT	$16.43_{(13.99,\ 18.86)}$	$4.89_{(4.43,\ 5.35)}$
			ST	$11.93_{(9.90,\ 13.95)}$	$3.69_{(3.26,\ 4.13)}$
			PST	$14.77_{(12.40,\ 17.13)}$	$3.67_{(3.25,\ 4.09)}$
			Eveno	$21.18_{(18.99,\ 23.38)}$	$9.13_{(9.08,\ 9.18)}$
		MI	*D*4, *ng*256	$38.87_{(36.07,\ 41.67)}$	$9.14_{(9.03,\ 9.25)}$
		CHMM	*D*4, *ng*4	$10.59_{(8.49,\ 12.69)}$	$2.93_{(2.53,\ 3.32}$

Column *Method* indicates the synchrony measurement method (WT, ST or PST) for correlation-based algorithms (CoIA and CANCOR) or the design parameters for the probabilistic algorithms (MI and CHMM), or the fusion algorithm used for fusion (sum rule or GMM fusion)

always results in a small (yet not statistically significant) improvement over the self-training approach.

MI This approach seems to be the least successful of all those tested in this paper for asynchrony detection. However, this does not mean that MI should not be used for monologue detection or speaker association. What it means is that the technique may not be appropriate when a global threshold is required, as in the case of a biometric application or a synchrony quality assessment task.

Sum rule and GMM fusion Performance improved when CoIA and CHMM were fused. The two systems encoded different types of synchrony data, and hence, when fused, resulted in improved performance, even though the two systems were using the same audiovisual features.

9 Conclusion and future work

The results reported in Sect. 7 demonstrate the effectiveness of both CoIA and CHMM as asynchrony detection methods. They have been tested in a difficult framework for asynchrony detection, where the video sequences and voice are taken from the same user uttering the same speech.

Asynchrony detection can be a useful anti-spoofing technique for real-life impostor attacks in biometric identity verification systems, among other applications such as speaker location and monologue detection.

The methods we presented can easily be adapted to identity verification systems based on audiovisual speech features. Client-dependent models can be derived, which

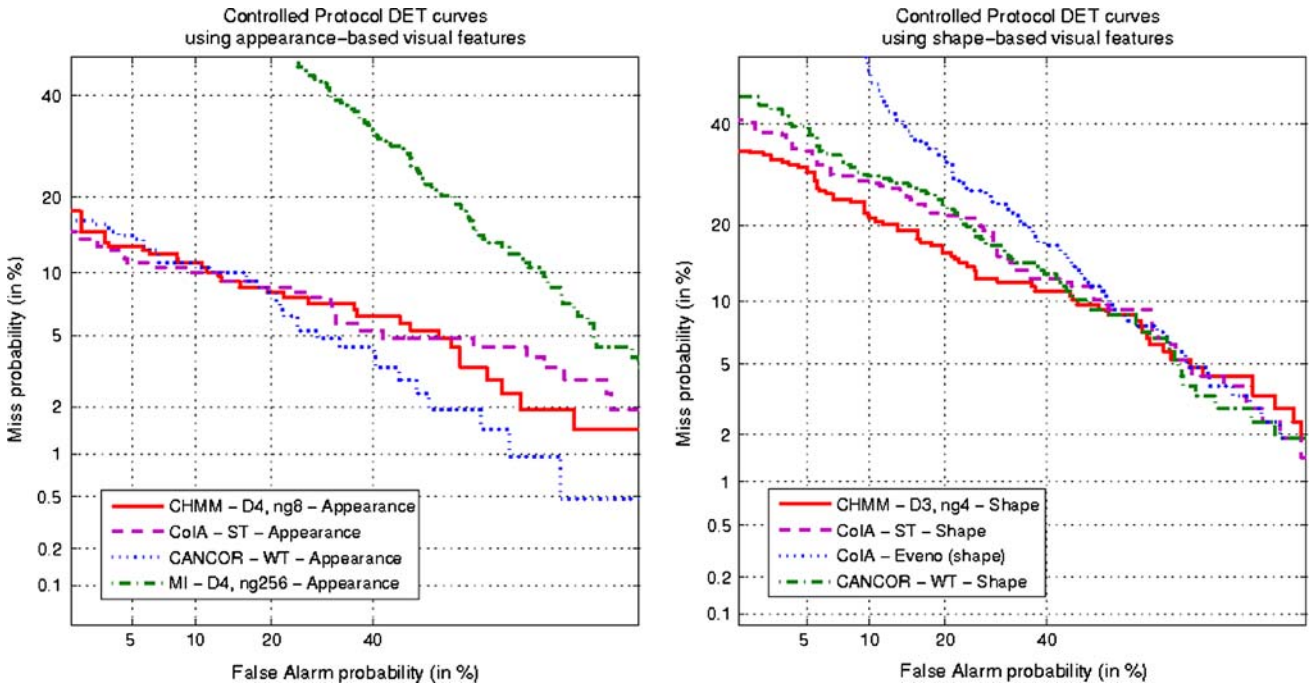


Fig. 9 Controlled protocol DET curves for the best methods shown in Table 1 using appearance-based (*left*) and shape-based visual features (*right*)

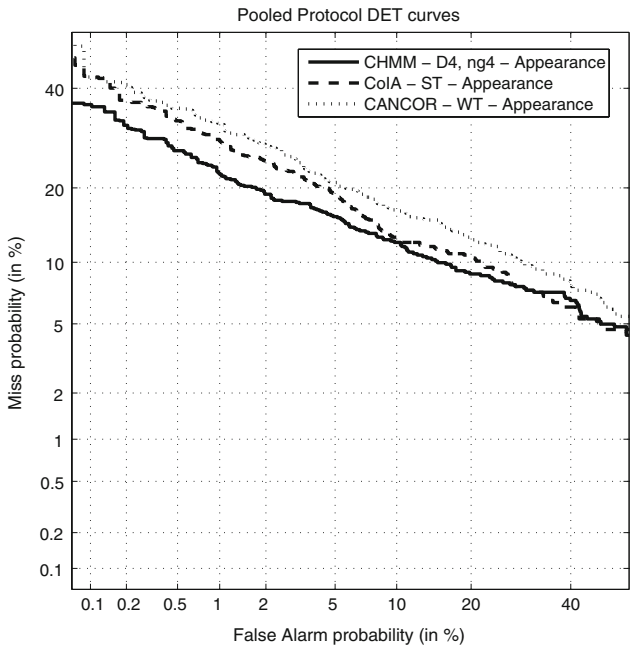


Fig. 10 Pooled protocol DET curves for the best methods shown in Table 1 using appearance-based and shape-based visual features

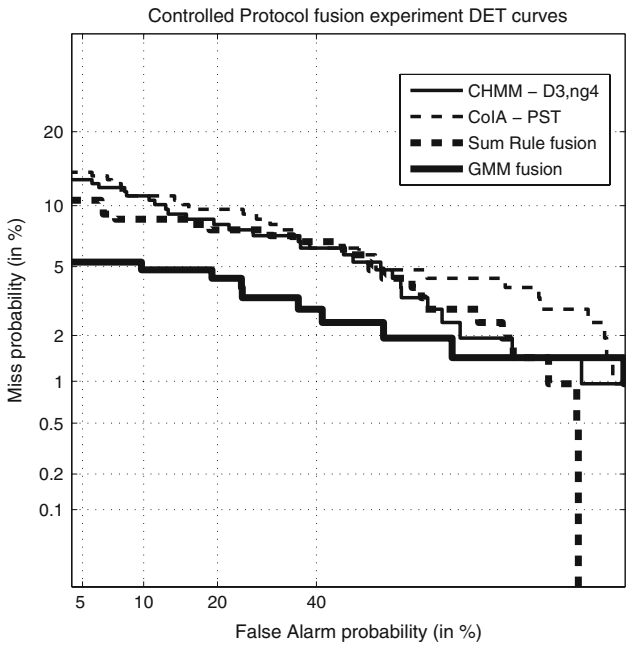


Fig. 11 Controlled protocol DET curves for the best CoIA and CHMM methods, sum rule and GMM fusion algorithms

would provide complementary information to speaker or face verifiers working in a multimodal framework.

Synchrony evaluation could also be used in other fields that are not directly related to biometrics, speaker location or monologue detection. It could be used, for example, to replace tasks that are currently done manually, such as the alignment of video and soundtrack in movie post-production, or the evaluation of the quality of dubbing into a foreign language.

New directions of research in asynchrony detection emerge from this paper. We have shown how fusing CoIA and CHMM scores can lead to improved

performance. Appearance- and shape-based systems can also be fused at feature level. The two systems offer different methods for integrating multiple information sources: while CoIA can be applied to concatenated appearance- and shape-based visual features, CHMM can work with an acoustic stream, an appearance-based visual feature stream and a shape-based visual feature stream. Structural improvements to CHMM are also a possibility in future studies given that CHMM families have already been used successfully for audiovisual speech recognition purposes [23, 24]. A large number of training audiovisual utterances containing different phonetic units is required if acceptable speech recognition accuracy is to be achieved. The uncoupling procedure described in Sect. 4.2 can be applied to such a CHMM audiovisual speech recognizer to obtain an asynchrony detector. The results would more than likely be much more accurate than those described here. Our system suffered from structural limitations due to insufficient training material and this resulted in poor audiovisual speech unit modeling, which is mostly based on the evolution of the most correlated components of both streams. Another promising research direction emerges from recently derived tensor based classification frameworks [25, 26]. Tensors encoding audio-visual speech features from several consecutive sampling periods can keep most of the dynamic relationship between lips movement and speech dynamics, while the use of tensors algebra can overcome the scarcity of training data. Equivalent CoIA and CANCOR tensor techniques should be derived in a future work and tested in the audio-visual asynchrony detection problem presented in this paper.

Acknowledgments This work has been partially supported by Spanish Ministry of Education and Science (project PRESA TEC2005-07212), by the Xunta de Galicia (project PGIDIT05TIC32202PR) and by the European Union through the European Networks of Excellence BioSecure and K-Space.

References

1. Potamianos G, Neti C, Luettin J, Matthews I (2004) Audio-visual automatic speech recognition: an overview. Issues Vis Audio Vis Speech Process
2. Liu X, Liang L, Zhaa Y, Pi X, Nefian AV (2002) Audio-visual continuous speech recognition using a coupled hidden Markov model. In: Proceedings of the international conference on spoken language processing
3. Gurbuz S, Tufekci Z, Patterson T, Gowdy JN (2002) Multi-stream product modal audio-visual integration strategy for robust adaptive speech recognition. In: Proceedings of IEEE international conference on acoustics, speech and signal processing, Orlando
4. Chibelushi CC, Deravi F, Mason JSD (2002) A review of speech-based bimodal recognition. IEEE Trans Multimed 4(1):23–37
5. Pan H, Liang Z-P, Huang TS (2000) A new approach to integrate audio and visual features of speech. In: IEEE international conference on multimedia and expo., pp 1093 – 1096
6. Chaudhari UV, Ramaswamy GN, Potamianos G, Neti C (2003) Information fusion and decision cascading for audio-visual speaker recognition based on time-varying stream reliability prediction. In: IEEE international conference on multimedia expo., vol III. Baltimore, pp 9–12, July 2003
7. Chetty G, Wagner M (2004) "Liveness" verification in audio-video authentication. In: Australian international conference on speech science and technology, pp 358–363
8. Eveno N, Besacier L (2005) A speaker independent liveness test for audio-video biometrics. In: Nineth European conference on speech communication and technology
9. Hershey J, Movellan J (2000) Audio vision: using audiovisual synchrony to locate sounds. In: Advances in neural information processing systems, vol 12, pp 813–819
10. Slaney M, Covell M (2000) FaceSync: a linear operator for measuring synchronization of video facial images and audio tracks. Neural Inf Process Soc 13
11. Fisher JW, Darell T (2004) Speaker association with signal-level audiovisual fusion. IEEE Trans Multimed 6(3):406–413
12. Nock HJ, Iyengar G, Neti C (2002) Assessing face and speech consistency for monologue detection in video. Multimedia 303–306
13. Bredin H, Chollet G (2006) Measuring audio and visual speech synchrony: methods and applications. In: International conference on visual information engineering
14. Lucas BD, Kanade T (1981) An iterative image registration technique with an application to stereo vision. In: DARPA image understanding workshop, pp 121–130
15. Bredin H, Aversano G, Mokbel C, Chollet G (2006) The biosecure talking-face reference system. In: Second workshop on multimodal user authentication, May 2006
16. Dolédec S, Chessel D (1994) Co-inertia analysis: an alternative method for studying species-environment relationships. Freshw Biol 31:277–294
17. Bailly-Baillière E, Bengio E, Bimbot F, Hamouz M, Kittler J, Mariéthoz J, Matas J, Messer K, Popovici V, Porée F, Ruiz B, Thiran J-P (2003) The BANCA database and evaluation protocol. In: Lecture notes in computer science, vol 2688, pp 625–638, January 2003
18. Gutiérrez J, Rouas J-L, André-Obrecht R (2004) Weighted loss functions to make risk-based language identification fused decisions. In: IEEE Computer Society (ed). Proceedings of the 17th international conference on pattern recognition (ICPR'04)
19. Qian J-Z, Ross A, Jain A (2001) Information fusion in biometrics. In: Proceedings of 3rd international conference on audio- and video-based person authentication (AVBPA), pp 354–359, Sweden, June 2001
20. Martin A, Doddington G, Kamm T, Ordowski M, Przybocki M (1997) The DET curve in assessment of detection task performance. In: European conference on speech communication and technology, pp 1895–1898
21. Bailly-Bailliére E, Bengio S, Bimbot F, Hamouz M, Kittler J, Marióthoz J, Matas J, Messer K, Popovici V, Porée F, Ruiz B, Thiran J-P (2003) The banca database and evaluation protocol
22. Bengio S, Mariéthoz J (2004) A statistical significance test for person authentication. ODYSSEY 2004—the speaker and language recognition workshop, pp 237–244
23. Zhang X, Mersereau RM, Clements M (2002) Bimodal fusion in audio-visual speech recognition, vol 1. In: IEEE 2002 international conference on image processing, pp 964–967, September 2002
24. Nefian AV, Liang L, Pi X, Xiaoxiang L, Mao C, Murphy K (2002) A coupled HMM for audio-visual speech recognition. In:

Proceedings of the international conference on acoustics speech and signal processing (ICASSP02), May 2002
25. Tao D, Li X, Hu W, Maybank S, Wu X (2007) Supervised tensor learning. knowledge and information systems, 13(1):1–42
26. Tao D, Li X, Wu X, Maybank SJ (2007) General tensor discriminant analysis and gabor features for gait recognition. IEEE Trans Pattern Anal Mach Intell 29(10):700–715

Author Biographies

Enrique Argones Rúa received the Telecommunications Engineer degree from the Universidad de Vigo, Spain, in 2003, where he is currently pursuing the Ph.D. degree in the field of audio-visual biometrics. His research interests include speaker verification, multimodal fusion and video-based face verification.

Hervé Bredin received his Ph.D. degree from the Signal and Image Processing Department of Télécom Paris Tech in 2007, focusing mostly on audiovisual identity verification based on talking-faces and its robustness to high-effort forgery (such as replay attacks, face animation or voice transformation). He is now with the Center for Digital Video Processing at Dublin City University.

Carmen García Mateo received the M.Sc. and Ph.D. degrees (Hons.) in telecommunications engineering from the Universidad Politécnicade Madrid, Spain, in 1987 and 1993, respectively. Her research interests include speech, and speaker recognition, dialogue systems and biometric applications. She has been the leader of a number of R&D projects and published papers on these topics. She is Professor in the field of Discrete Signal processing at the Universidad de Vigo.

Gérard Chollet studied Linguistics, Electrical Engineering and Computer Science at the University of California, Santa Barbara where he was granted a Ph.D. in Computer Science and Linguistics. He joined CNRS in 1978 at the Institut de Phonétique in Aix en Provence. In 1992, he participated to the development of IDIAP, a new research laboratory of the "Fondation Dalle Molle" in Martigny, Switzerland. Since 1996, he is at ENST, managing research projects and supervising doctoral work. His main research interests are in phonetics, automatic speech processing, speech dialog systems, multimedia, pattern recognition, digital signal processing, speech pathology, speech training aids,…

Daniel González Jiménez received the Telecommunications Engineer degree from the Universidad de Vigo, Spain, in 2003, where he is currently pursuing the Ph.D. degree in the field of face-based biometrics. His research interests include computer vision and image processing.