Hindawi Publishing Corporation
EURASIP Journal on Advances in Signal Processing
Volume 2007, Article ID 70186, 11 pages
doi:10.1155/2007/70186

Research Article

Audiovisual Speech Synchrony Measure: Application to Biometrics

Hervé Bredin and Gérard Chollet

Département Traitement du Signal et de l'Image, École Nationale Supérieure des Télécommunications, CNRS/LTCI, 46 rue Barrault, 75013 Paris Cedex 13, France

Received 18 August 2006; Accepted 18 March 2007

Recommended by Ebroul Izquierdo

Speech is a means of communication which is intrinsically bimodal: the audio signal originates from the dynamics of the articulators. This paper reviews recent works in the field of audiovisual speech, and more specifically techniques developed to measure the level of correspondence between audio and visual speech. It overviews the most common audio and visual speech front-end processing, transformations performed on audio, visual, or joint audiovisual feature spaces, and the actual measure of correspondence between audio and visual speech. Finally, the use of synchrony measure for biometric identity verification based on talking faces is experimented on the BANCA database.

1. INTRODUCTION

Speech is a means of communication which is intrinsically bimodal: the audio signal originates from the dynamics of the articulators. Both audible and visible speech cues carry relevant information. Though the first automatic speech-based recognition systems were only relying on its auditory part (whether it is speech recognition or speaker verification), it is well known that its visual counterpart can be a great help, especially under adverse conditions [1]. In noisy environments for example, audiovisual speech recognizers perform better than audio-only systems. Using visual speech as a second source of information for speaker verification has also been experimented, even though resulting improvements are not always significant.

This review tries to complement existing surveys about audiovisual speech processing. It does not address the problem of audiovisual speech recognition nor speaker verification: these two issues are already covered in [2, 3]. Moreover, this paper does not tackle the question of the estimation of visual speech from its acoustic counterpart (or reciprocally): the reader might want to have a look at [4, 5] showing that linear methods can lead to very good estimates.

This paper focuses on the measure of correspondence between acoustic and visual speech. How correlated the two signals are? Can we detect a lack of correspondence between them? Is it possible to decide (putting aside any biometric method), among a few people appearing in a video, who is talking?

Section 2 overviews the acoustic and visual front-ends processing. They are often very similar to the one used for speech recognition and speaker verification, though a tendency to simplify them as much as possible has been noticed. Moreover, linear transformations aiming at improving joint audiovisual modeling are often performed as a preliminary step before measuring the audiovisual correspondence, they will be discussed in Section 3. The correspondence measures proposed in the literature are then presented in Section 4. The results that we obtained in the biometric identity verification task using synchrony measures on the BANCA [6] database are presented in Section 5. Finally, a list of other applications of these techniques in different technological areas is presented in Section 6.

2. FRONT-END PROCESSING

This section reviews the speech front-end processing techniques used in the literature for audiovisual speech processing in the specific framework of audiovisual speech synchrony measures. They all share the common goal of reducing the raw data in order to achieve a good subsequent modeling.

2.1. Acoustic speech processing

Acoustic speech parameterization is classically performed on overlapping sliding windows of the original audio signal.

Short-time energy

The raw amplitude of the audio signal can be used as is. In [7], the authors extract the average acoustic energy on the current window as their one-dimensional audio feature. Similar methods such as root mean square amplitude or log-energy were also proposed [4, 8].

Periodogram

In [9], a [0–10 kHz] periodogram of the audio signal is computed on a sliding window of length 2/29.97 seconds (corresponding to the duration of 2 frames of the video) and directly used as the parameterization of the audio stream.

Mel-frequency cepstral coefficients

The use of MFCC parameterization is very frequent in the literature [10–14]. There is a practical reason for that; it is the state-of-the-art [15] parameterization for speech processing in general, including speech recognition and speaker verification.

Linear predictive coding and line spectral frequencies

Linear predictive coding, and its derivation line spectral frequencies [16], have also been widely investigated. The latter are often preferred because they are directly related to the vocal tract resonances [5].

A comparison of these different acoustic speech features is performed in [14] in the framework of the *FaceSync* linear operator, which is presented in Section 3.3. To summarize, in their specific framework, the authors conclude that MFCC, LSF and LPC parameterizations lead to a stronger correlation with the visual speech than spectrogram and raw energy features. This result is coherent with the observation that these features are the ones known to give good results for speech recognition.

2.2. Visual speech processing

In this section, we will refer to the gray-level mouth area as the region of interest. It can be much larger than the sole lip area and can include jaw and cheeks. In the following, it is assumed that the detection of this region of interest has already been performed. Most of visual speech features proposed in the literature are shared by studies in audiovisual speech recognition. However, some much more simple visual features are also used for synchronization detection.

Raw intensity of pixels

This is the visual equivalent of the audio raw energy. In [7, 12], the intensity of gray-level pixels is used as is. In [8], their sum over the whole region of interest is computed, leading to a one-dimensional feature.

Holistic methods

Holistic methods consider and process the region of interest as a whole source of information. In [13], a two-dimensional discrete cosine transform (DCT) is applied on the region of interest and the most energetic coefficients are kept as visual features, it is a well-known method in the field of image compression. Linear transformations taking into account the specific distribution of gray-level in the region of interest were also investigated. Thus, in [17], the authors perform a projection of the region of interest on vectors resulting from a principal component analysis; they call the principal components "eigenlips" by analogy with the well-known "eigenfaces" [18] principle used for face recognition.

Lip-shape methods

Lip-shape methods consider and process lips as a deformable object from which geometrical features can be derived, such as height, width openness of the mouth, position of lip corners, and so forth. They are often based on fiducial points which need to be automatically located. In [4], videos available are recorded using two cameras (one frontal, one from side) and the automatic localization is made easier by the use of face make-up, both frontal and profile measures are then extracted and used as visual features. Mouth width, mouth height, and lip protrusion are computed in [19], jointly with what the authors call the relative teeth count which can be considered as a measure of the visibility of teeth. In [20, 21], a deformable template composed of several polynomial curves follows the lip contours; it allows the computation of the mouth width, height, and area. In [10], the lip shape is summarized with a one-dimensional feature, the ratio of lip height and lip width.

Dynamic features

In [3], the authors underline that, though it is widely agreed that an important part of speech information is conveyed dynamically, dynamic features extraction is rarely performed; this observation is also verified for correspondence measures. However, some attempts to capture dynamic information within the extracted features do exist in the literature. Thus, the use of time derivatives is investigated in [22]. In [11], the authors compute the total temporal variation (between two subsequent frames) of pixel values in the region of interest, following: (1)

$$v_t = \sum_{x=1}^{W} \sum_{y=1}^{H} \left| I_t(x, y) - I_{t+1}(x, y) \right|, \tag{1}$$

where $I_t(x, y)$ is the grey-level pixel value of the region of interest at position (x, y) in frame t.

2.3. Frame rates

Audio and visual sample rates are classically very different. For speaker verification, for example, MFCCs are usually extracted every 10 milliseconds; whereas videos are often encoded at a frame rate of 25 images per second. Therefore, it is often required to downsample audio features or upsample visual features in order to equalize audio and visual sample rates. However, though the extraction of raw energy or periodogram can be performed directly on a larger window, downsampling audio features is known to be very bad for speech recognition. Therefore, upsampling visual features is often preferred (with linear interpolation, e.g.). One could also think of using a camera able to produce 100 images per second. Finally, some studies (like the one presented in Section 4.3.2) directly work on audio and visual features with unbalanced sample rates.

3. AUDIOVISUAL SUBSPACES

In this section, we overview transformations that can be applied on audio, visual, and/or audiovisual spaces with the aim of improving subsequent measure of correspondence between audio and visual clues.

3.1. Principal component analysis

Principal component analysis (PCA) is a well-known linear transformation which is optimal for keeping the subspace that has largest variance. The basis of the resulting subspace is a collection of principal components. The first principal component corresponds to the direction of the greatest variance of a given dataset. The second principal component corresponds to the direction of second greatest variance, and so on. In [23], PCA is used in order to reduce the dimensionality of a joint audiovisual space (in which audio speech features and visual speech features are concatenated) while keeping the characteristics that contribute most to its variance.

3.2. Independent component analysis

Independent component analysis (ICA) was originally introduced to deal with the issue of source separation [24]. In [25], the authors use visual speech features to improve separation of speech sources. In [26], ICA is applied on an audiovisual recording of a piano session, and a close-up of the keyboard is shot when the microphone is recording the music. ICA allows to clearly find a correspondence between the audio and visual note. However, to our knowledge, ICA has never been used as a transformation of the audiovisual speech feature space (as in [26] for the piano). A Matlab implementation of ICA is available on the Internet [27].

3.3. Canonical correlation analysis

Canonical correlation analysis (CANCOR) is a multivariate statistical analysis allowing to jointly transform the audio and visual feature spaces while maximizing their correlation in the resulting transformed audio and visual feature spaces. Given two synchronized random variables X and Y, the *FaceSync* algorithm presented in [14] uses CANCOR to find canonic correlation matrices $\mathbf{A}$ and $\mathbf{B}$ that whiten X and Y under the constraint of making their cross-correlation diagonal and maximally compact. Let $\mathcal{X} = (X - \mu_X)^T\mathbf{A}$, $\mathcal{Y} = (Y - \mu_Y)^T\mathbf{B}$, and $\Sigma_{\mathcal{X}\mathcal{Y}} = \mathbb{E}[\mathcal{X}\mathcal{Y}^T]$. These constraints can be summarized as follows:

whitening: $\mathbb{E}[\mathcal{X}\mathcal{X}^T] = \mathbb{E}[\mathcal{Y}\mathcal{Y}^T] = I$,

diagonal: $\Sigma_{\mathcal{X}\mathcal{Y}} = \mathrm{diag}\{\sigma_1, \ldots, \sigma_M\}$ with $1 \geq \sigma_1 \geq \cdots \geq \sigma_m > 0$ and $\sigma_{m+1} = \cdots = \sigma_M = 0$,

maximally compact: for i from 1 to M, the correlation $\sigma_i = \mathrm{corr}(\mathcal{X}_i, \mathcal{Y}_i)$ between $\mathcal{X}_i$ and $\mathcal{Y}_i$ is as large as possible.

The proof of the algorithm for computing $\mathbf{A} = [\mathbf{a}_1, \ldots, \mathbf{a}_m]$ and $\mathbf{B} = [\mathbf{b}_1, \ldots, \mathbf{b}_m]$ is described in [14]. One can show that the $\mathbf{a}_i$ are the normalized eigenvectors (sorted in decreasing order of their corresponding eigenvalue) of matrix $C_{XX}^{-1}C_{XY}C_{YY}^{-1}C_{YX}$ and $\mathbf{b}_i$ is the normalized vector which is collinear to $C_{YY}^{-1}C_{YX}\mathbf{a}_i$, where $C_{XY} = \mathrm{cov}(X, Y)$. A Matlab implementation of this transformation is also available on the Internet [28].

3.4. Coinertia analysis

Coinertia analysis (CoIA) is quite similar to CANCOR. However, while CANCOR is based on the maximization of the correlation between audio and visual features, CoIA relies on the maximization of their covariance $\mathrm{cov}(\mathcal{X}_i, \mathcal{Y}_i) = \mathrm{corr}(\mathcal{X}_i, \mathcal{Y}_i) \times \mathrm{var}(\mathcal{X}_i) \times \mathrm{var}(\mathcal{Y}_i)$. This statistical analysis was first introduced in biology and is relatively new in our domain. The proof of the algorithm for computing $\mathbf{A}$ and $\mathbf{B}$ can be found in [29]. One can show that the $\mathbf{a}_i$ are the normalized eigenvectors (sorted in decreasing order of their corresponding eigenvalue) of matrix $C_{XY}C_{XY}^t$ and $\mathbf{b}_i$ is the normalized vector which is collinear to $C_{XY}^t\mathbf{a}_i$.

Remark 1. Comparative studies between CANCOR and CoIA are proposed in [19–21]. The authors of [19] show that CoIA is more stable than CANCOR; the accuracy of the results is much less sensitive to the number of samples available. The *liveness* score (see Section 6) proposed in [20, 21] is much more efficient with CoIA than CANCOR. The authors of [21] suggest that this difference is explained by the fact that CoIA is a compromise between CANCOR (where audiovisual correlation is maximized) and PCA (where audio and visual variances are maximized) and therefore benefits from the advantages of both transformations.

4. CORRESPONDENCE MEASURES

This section overviews the correspondence measures proposed in the literature to evaluate the synchrony between audio and visual features resulting from audiovisual front-end processing and transformations described in Sections 2 and 3.

4.1. Pearson's product-moment coefficient

Let X and Y be two independent random variables which are normally distributed. The square of their Pearson's product-moment coefficient $R(X, Y)$ (defined in (2)) denotes the portion of total variance of X that can be explained by a linear transformation of Y (and reciprocally, since it is a symmetrical measure):

$$R(X, Y) = \frac{\operatorname{cov}(X, Y)}{\sigma_X \sigma_Y}. \tag{2}$$

In [7], the authors compute the Pearson's product-moment coefficient between the average acoustic energy X and the value Y of the pixels of the video to determine which area of the video is more correlated with the audio. This allows to decide which of two people appearing in a video is talking.

4.2. Mutual information

In information theory, the mutual information $MI(X, Y)$ of two random variables X and Y is a quantity that measures the mutual dependence of the two variables. In the case of X and Y are discrete random variables, it is defined as in (3),

$$MI(X, Y) = \sum_{x \in X} \sum_{y \in Y} p(x, y) \log \frac{p(x, y)}{p(x)p(y)}. \tag{3}$$

It is nonnegative ($MI(X, Y) \geq 0$) and symmetrical ($MI(X, Y) = MI(Y, X)$). One can demonstrate that X and Y are independent if and only if $MI(X, Y) = 0$. The mutual information can also be linked to the concept of entropy H in information theory as shown in (5):

$$MI(X, Y) = H(X) - H(X \mid Y), \tag{4}$$

$$MI(X, Y) = H(X) + H(Y) - H(X, Y). \tag{5}$$

As shown in [7], in the special case where X and Y are normally distributed monodimensional random variables, the mutual information is related to $R(X, Y)$ via the following equation:

$$MI(X, Y) = -\frac{1}{2} \log \left(1 - R(X, Y)^2\right). \tag{6}$$

In [7, 12, 13, 30], the mutual information is used to locate the pixels in the video which are most likely to correspond to the audio signal, the face of the person who is speaking clearly corresponds to these pixels. However, one can notice that the mouth area is not always the part of the face with the maximum mutual information with the audio signal, it is very dependent on the speaker.

Remark 2. In [17], the mutual information between audio X and time-shifted visual Y_t features is plotted as a function of their temporal offset t. It shows that the mutual information reaches its maximum for a visual delay of between 0 and 120 milliseconds. This observation led the authors of [20, 21] to propose a liveness score $L(X, Y)$ based on the maximum value R_{ref} of the Pearson's coefficient for short time offset between audio and visual features,

$$R_{\text{ref}} = \max_{-2 \leq t \leq 0} \left[R(X, Y_t)\right]. \tag{7}$$

4.3. Joint audiovisual models

Though the Pearson's coefficient and the mutual information are good at measuring correspondence between two random variables even if they are not linearly correlated (which is what they were primarily defined for), some other methods does not rely on this linear assumption.

4.3.1. Parametric models

Gaussian mixture models

Let us consider two discrete random variables $X = \{x_t, t \in \mathbb{N}\}$ and $Y = \{y_t, t \in \mathbb{N}\}$ of dimension d_X and d_Y, respectively. Typically, X would be acoustic speech features and Y visual speech features [10, 31]. One can define the discrete random variable $Z = \{z_t, t \in \mathbb{N}\}$ of dimension d_Z where z_t is the concatenation of the two samples x_t and y_t, such as $z_t = [x_t, y_t]$ and $d_Z = d_X + d_Y$.

Given a sample z, the Gaussian mixture model λ defines its probability distribution function as follows:

$$p(z \mid \lambda) = \sum_{i=1}^{N} w_i \mathcal{N}(z; \mu_i, \Gamma_i), \tag{8}$$

where $\mathcal{N}(\bullet; \mu, \Gamma)$ is the normal distribution of mean μ and covariance matrix $\Gamma \cdot \lambda = \{w_i, \mu_i, \Gamma_i\}_{i \in [1,N]}$ are parameters describing the joint distribution of X and Y. Using a training set of synchronized samples x_t and y_t concatenated into joint samples z_t, the Expectation-Maximization algorithm (EM) allows the estimation of λ.

Given two sequences of test $X = \{x_t, t \in [1, T]\}$ and $Y = \{y_t, t \in [1, T]\}$, a measure of their correspondence $C_\lambda(X, Y)$ can be computed as in (9),

$$C_\lambda(X, Y) = \frac{1}{T} \sum_{t=1}^{T} p([x_t, y_t] \mid \lambda). \tag{9}$$

Then the application of a threshold θ decides on whether the acoustic speech X and the visual speech Y correspond to each other (if $C_\lambda(X, Y) > \theta$) or not (if $C_\lambda(X, Y) \leq \theta$).

Remark 3. λ is well known to be speaker-dependent, GMM-based systems are the state-of-the-art for speaker identification. However, there is often not enough training samples from a speaker S to correctly estimate the model λ_S using the EM algorithm. Therefore, one can adapt a world model λ_Ω (estimated on a large set of training samples from a population as large as possible) using the few samples available from speaker S into a model λ_S. This is not the purpose of this paper to review adaptation techniques, the reader can refer to [15] for more information.

Hidden Markov models

Like the Pearson's coefficient and the mutual information, time offset between acoustic and visual speech features is not modeled using GMMs. Therefore, the authors of [13] propose to model audiovisual speech with hidden Markov

models (HMMs). Two speech recognizers are trained: one classical audio only recognizer [32], and an audiovisual speech recognizer as described in [1]. Given a sequence of audiovisual samples ($[x_t, y_t], t \in [1, T]$), the audio-only system gives a word hypothesis W. Then, using the HMM of the audiovisual system, what the authors call a measure of plausibility $P(X, Y)$ is computed as follows:

$$P(X, Y) = p([x_1, y_1] \cdots [x_T, y_T] \mid W). \tag{10}$$

An asynchronous hidden Markov model (AHMM) for audiovisual speech recognition is proposed in [33]. It assumes that there is always an audio observation x_t and sometimes a visual observation y_s at time t. It intrinsically models the difference of sample rates between audio and visual speech, by introducing the probability that the system emits the next visual observation y_s at time t. AHMM appears to outperform HMM in the task of audiovisual speech recognition [33] while naturally resolving the problem of different audio and visual sample rates.

4.3.2. Nonparametric models

The use of neural networks (NN) is investigated in [11]. Given a training set of both synchronized and not synchronized audio and visual speech features, a neural network with one hidden layer is trained to output 1 when the audiovisual input features are synchronized and 0 when they are not. Moreover, the authors propose to use an input layer at time t consisting of $[X_{t-N_X}, \ldots, X_t, \ldots, X_{t+N_X}]$ and $[Y_{t-N_Y}, \ldots, Y_t, \ldots, Y_{t+N_Y}]$ (instead of X_t and Y_t), choosing N_X and N_Y such as about 200 milliseconds of temporal context is given as an input. This proposition is a way of solving the well-known problem of coarticulation and the already mentioned lag between audio and visual speech. It also removes the need for down-sampling audio features (or upsampling visual features).

5. APPLICATION TO BIOMETRICS

Among many applications (some of which are listed in Section 6), identity verification based on talking faces is one that can really benefit from synchrony measures.

5.1. Audiovisual features extraction

Given an audiovisual sequence AV, we use our algorithm for face and lip tracking [34] to locate the lip area in every frame, as shown in Figure 1. While 15 classical MFCC coefficients are extracted every 10 milliseconds from the audio of the sequence AV, the first 30 DCT coefficients of the grey-level lip area are extracted (in a zigzag manner) from every frame of the video. A linear interpolation is finally performed on the visual features to reach the audio sample rate (100 Hz). This feature extraction process is done for every sequence AV to get the two random variables $X \in \mathbb{R}^{15}$ (for audio speech) and $Y \in \mathbb{R}^{30}$ (for visual speech).

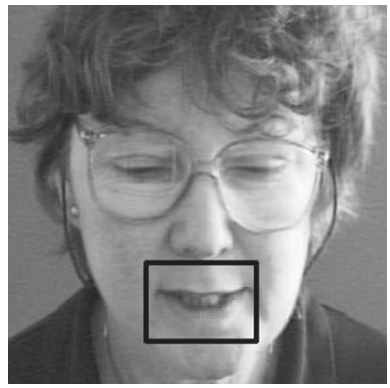

Figure 1: Lip tracking on the BANCA database.

5.2. Synchrony measures

We introduce two novel synchrony measures $\dot{S}$ and $\ddot{S}$ based on Canonical correlation analysis and Co-inertia analysis, respectively. The first step is to compute the transformation matrices $\dot{\mathbf{A}}$, and $\dot{\mathbf{B}}$ for CCA (resp., $\ddot{\mathbf{A}}$ and $\ddot{\mathbf{B}}$ for CoIA). A training set made of a collection of synchronized audiovisual sequences is gathered to compute them, using the formulae described in [29] (resp., in [14]). Consequently, we can define the following audiovisual speech synchrony measures in (11) and (12):

$$\dot{S}_{\dot{\mathbf{A}},\dot{\mathbf{B}}}(X, Y) = \frac{1}{K} \sum_{k=1}^{K} \left| \operatorname{corr}(\dot{a}_k^T X, \dot{b}_k^T Y) \right| \tag{11}$$

$$\ddot{S}_{\ddot{\mathbf{A}},\ddot{\mathbf{B}}}(X, Y) = \frac{1}{K} \sum_{k=1}^{K} \left| \operatorname{cov}(\ddot{a}_k^T X, \ddot{b}_k^T Y) \right|, \tag{12}$$

where only the first K vectors a_k and b_k of matrices $\mathbf{A}$ and $\mathbf{B}$ are considered. In the following, we will arbitrarily choose $K = 3$.

5.3. Replay attacks

Most of audiovisual identity verification systems based on talking faces perform a fusion of the scores given by a speaker verification algorithm and a face recognition algorithm. Therefore, it is quite easy for an impostor to impersonate his/her target if he/she owns recordings of his/her voice and pictures (or videos) of his/her face.

5.3.1. Impersonation scenarios

Many databases are available to the research community to help evaluate multimodal biometric verification algorithms, such as BANCA [6], XM2VTS [35], BT-DAVID [36], BIOMET [37], MyIdea, and IV2. Different protocols have been defined for evaluating biometric systems on each of these databases, but they share the assumption that impostor attacks are *zero-effort* attacks, that is, that the impostors use their own voice and face to perform the impersonation trial. These attacks are of course quite unrealistic, only a fool

would attempt to imitate a person without knowing anything about them.

Therefore, in [8], we have augmented the original BANCA protocols with more realistic impersonation scenarios, which can be divided into two categories: forgery scenarios (where voice and/or face transformation is performed) and replay attacks scenarios (where previously acquired biometric samples are used to impersonate the target).

In this section, we will tackle the *Big Brother* scenario; prior to the attack, the impostor records a movie of the target's face and acquires a recording of his/her voice. However, the audio and video do not come from the same utterance, so they may not be synchronized. This is a realistic assumption in situations where the identity verification protocol chooses an utterance for the client to speak.

5.3.2. Training

As mentioned earlier, a preliminary training step is needed to learn the projection matrices **A** and **B** (both for CCA and CoIA) and—then only—the synchrony measures can be computed. This training step can be done using different training sets depending on the targeted application.

World model

In this configuration, a large training set of synchronized audiovisual sequences is used to learn **A** and **B**.

Client model

The use of a client-dependent training set (of synchronized audiovisual sequences from one particular person) will be more deeply investigated in Section 5.4 about identity verification.

No training

One could also avoid the preliminary training set by learning (at test time) **A** and **B** on the tested audiovisual sequence (X, Y) itself.

Self-training

This method is an improvement brought to the above and was driven by the following intuition: *It is possible to learn a synchrony model between synchronized variables, whereas nothing can be learned from not-synchronized variables.* Given a tested audiovisual sequence (X, Y), with $X = \{\mathbf{x_1}, \ldots, \mathbf{x_N}\}$ and $Y = \{\mathbf{y_1}, \ldots, \mathbf{y_N}\}$, one can therefore try to learn the projection matrices **A** and **B** from a subsequence $(X_{\text{train}} = \{\mathbf{x_1}, \ldots, \mathbf{x_L}\}, Y_{\text{train}} = \{\mathbf{y_1}, \ldots, \mathbf{y_L}\})$, with $L < N$ and compute the synchrony measure S on what is left of the sequence: $(X_{\text{test}}, Y_{\text{test}})$ with $X_{\text{test}} = \{\mathbf{x_{L+1}}, \ldots, \mathbf{x_N}\}$ and $Y_{\text{test}} = \{\mathbf{y_{L+1}}, \ldots, \mathbf{y_N}\}$. In order to improve the robustness of this method, a cross-validation principle is applied: the partition between training and test set is performed P times by randomly drawing samples from (X, Y) to build the training set (keeping the others for the test set). Each partition p leads to a measure S_p and the final synchrony measure S is computed as their mean, $S = (1/P) \sum_{p=1}^{P} S_p$.

5.3.3. Experiments

Experiments are performed on the BANCA database [6], which is divided into two disjoint groups (G1 and G2) of 26 people. Each person recorded 12 videos where he/she says his/her own text (always the same) and 12 other videos where he/she says the text of another person from the same group, this makes 624 synchronized audiovisual sequences per group. On the other side, for each group, 14352 not-synchronized audiovisual sequences were artificially recomposed from audio and video from two different original sequences with one strong constraint that the person heard and the person seen pronounce the same utterance (in order to make the boundary decision between synchronized and not-synchronized audiovisual sequences even more difficult to define).

For each synchronized and not-synchronized sequence, a synchrony measure S is computed. This measure is then compared to a threshold θ and the sequence is decided to be synchronized if it is bigger than θ and not-synchronized otherwise. Varying the threshold θ, a DET curve [38] can be plotted. On the x-axis, the percentage of falsely rejected synchronized sequences is plotted, whereas the y-axis shows the percentage of falsely accepted not-synchronized sequences (depending on the chosen value for θ).

5.3.4. Results

Figure 2 shows the performance of the CCA (left) and CoIA (right) measures using the different training procedures described in Section 5.3.2. The best performance is achieved with the novel *Self-training* we introduced, both for CCA and CoIA, as well as with the CCA using *World model*, it gives an equal error rate (EER) of around 17%. It is noticeable that *World model* works better with CCA whereas *Client model* gives poor results with CCA and works nearly as good as *Self-training* with CoIA. This latter observation confirms what was previously noticed in [19]. The CoIA is much less sensitive to the number of training samples available, the CoIA works fine with little data (*Client model* only uses one BANCA sequence to train **A** and **B** [6]) and the CCA needs a lot of data for robust training.

Finally, Figure 3 shows that one can improve the performance of the algorithm for synchrony detection by fusing two scores (based on CCA and based on CoIA). After a classical step of score normalization, a support vector machine (SVM) with linear kernel is trained on one group (G1 or G2) and applied on the other one. The fusion of CCA with *World model* and CoIA with *Self-training* lowers the EER to around 14%. This final EER is comparable to what was achieved in [21].

5.4. Identity verification

According to the results obtained in Figure 2, not only can synchrony measures be used as a first barrier against replay

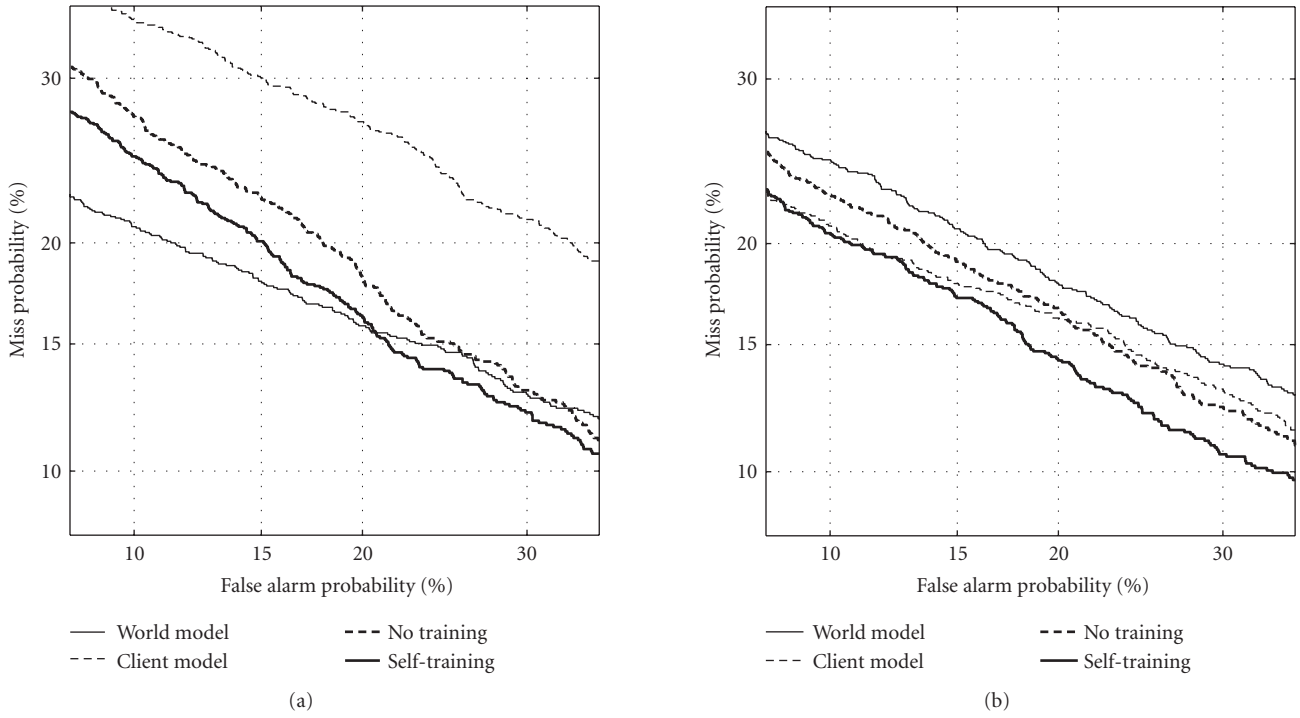

Figure 2: Synchrony detection with CCA and CoIA.

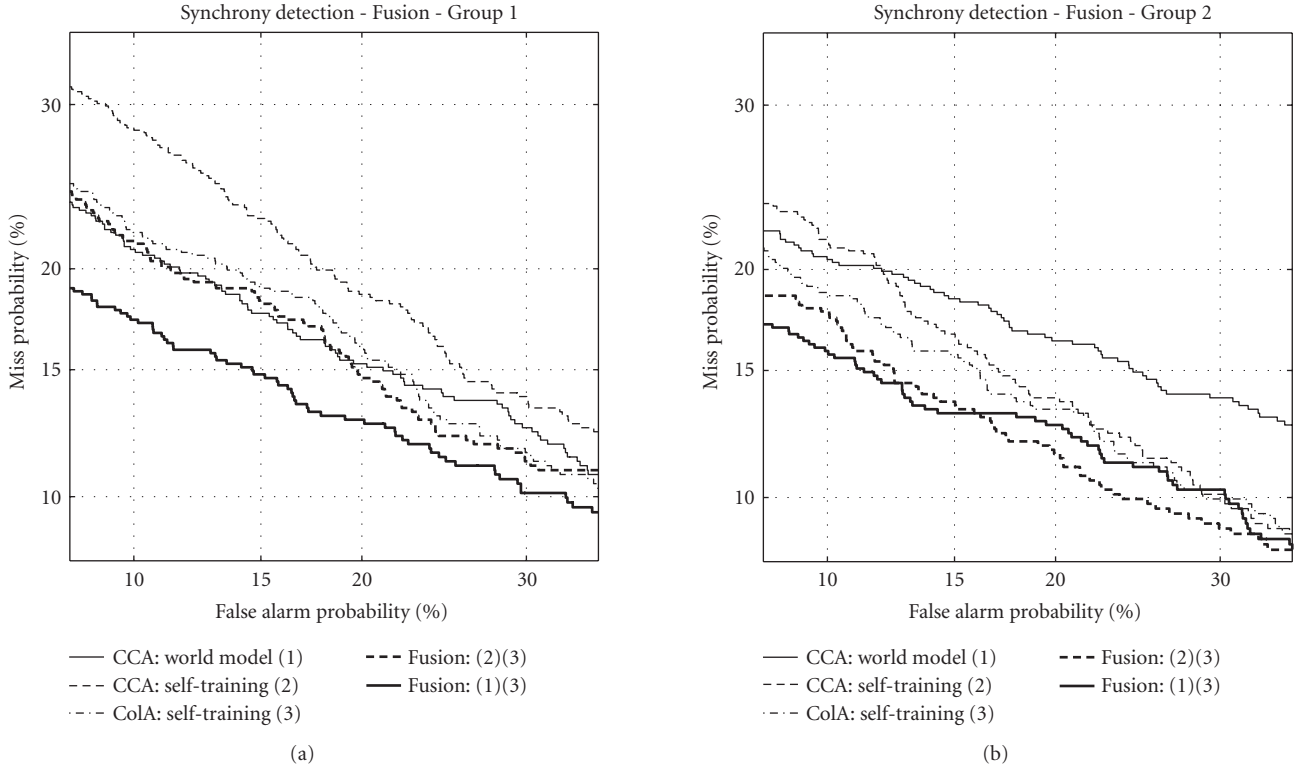

Figure 3: Fusion of CoIA and CCA.

attacks, but it also led us to investigate the use of audiovisual speech synchrony measure for identity verification (see performance achieved by the CoIA with *Client model*).

Some previous work have been done in identity verification using fusion of speech and lip motion. In [23] the authors apply classical linear transformations for dimensionality reduction (such as principal component analysis - PCA, or linear discriminant analysis—LDA) on feature vectors resulting from the concatenation of audio and visual speech features. CCA is used in [39] where projected audio and visual speech features are used as input for client-dependent HMM models.

Our novel approach uses CoIA with *Client model* (that achieved very good results for synchrony detection) to identify people with their personal way of synchronizing their audio and visual speech.

5.4.1. Principle

Given an enrollment audiovisual sequence AV_λ from a person λ, one can extract the corresponding synchronized variables X_λ and Y_λ as described in Section 5.2. Then, using (X_λ, Y_λ) as the training set, client-dependent CoIA projection matrices $\ddot{\mathbf{A}}_{\boldsymbol{\lambda}}$ and $\ddot{\mathbf{B}}_{\boldsymbol{\lambda}}$ are computed and stored as the model of client λ.

At test time, given an audiovisual sequence AV_ϵ from a person ϵ pretending to be the client λ, one can extract the corresponding variables X_ϵ and Y_ϵ. $\ddot{S}_{\ddot{\mathbf{A}}_\lambda, \ddot{\mathbf{B}}_\lambda}(X_\epsilon, Y_\epsilon)$ (defined in (12)) finally allows to get a score which can be compared to a threshold θ. The person ϵ is accepted as the client λ if $\ddot{S}_{\ddot{\mathbf{A}}_\lambda, \ddot{\mathbf{B}}_\lambda}(X_\epsilon, Y_\epsilon) > \theta$ and rejected otherwise.

5.4.2. Experiments

Experiments are performed on the BANCA database following the *Pooled* protocol [6]. The client access of the first session of each client is used as the enrollment data and the test are performed using all the other sequences (11 client accesses and 12 impostor accesses per person). The impostor accesses are *zero-effort* impersonation attacks since the impostor uses his/her own face and voice when pretending to be his/her target. Therefore, we also investigated replay attacks. The client accesses of the Pooled protocol are not modified, only the impostor accesses are, to simulate replay attacks.

Video replay attack

A video of the target is shown while the original voice of the impostor is kept unchanged.

Audio replay attack

The voice of the target is played while the original face of the impostor is kept unchanged.

Notice that, even though the acoustic and visual speech signals are not synchronized, the same utterance (a digit code and the name and address of the claimed identity) is pronounced.

5.4.3. Results

Figure 4 shows the performance of identity verification using the client-dependent synchrony model on these three protocols.

On the original *zero-effort* Pooled protocol, the algorithm achieves an EER of 32%. This relatively weak method might however bring some extra discriminative power to a system based only on the speech and face modalities, which we will study in the the following section. We can also notice that it is intrinsically robust to replay attacks: both audio and video replay attacks protocols lead to an EER of around 17%. This latter observation also shows that this new modality is very little correlated to the speech and face modality, and mostly depends on the actual correlation for which it was originally designed.

6. OTHER APPLICATIONS

Measuring the synchrony between audio and visual speech features can be a great help in many other applications dealing with audiovisual sequences.

Sound source localization

Sound source localization is the most cited application of audio and visual speech correspondence measure. In [11], a sliding window performs a scan of the video, looking for the most probable mouth area corresponding to the audio track (using a time-delayed neural network). In [13], the principle of mutual information allows to choose which of the four faces appearing in the video is the source of the audio track, the authors announce a 82% accuracy (averaged on 1016 video tests). One can think of an intelligent videoconferencing system making extensive use of such results, the camera could zoom in on the person who is currently speaking.

Indexation of audiovisual sequences

Another field of interest is the indexation of audiovisual sequences. In [12], the authors combine scores from three systems (face detection, speech detection, and a measure of correspondence based on the mutual information between the soundtrack and the value of pixels) to improve their algorithm for detection of monologue. Experiments performed in the framework of the TREC 2002 video retrieval track [40] show a 50% relative improvement on the average precision.

Film postproduction

During the postproduction of a film, dialogues are often rerecorded in a studio. An audiovisual speech correspondence measure can be of great help when synchronizing the new audio recording with the original video. Such measures can also be a way of evaluating the quality of a dubbed film into a foreign language: does the translation fit well with the original actor facial motions?

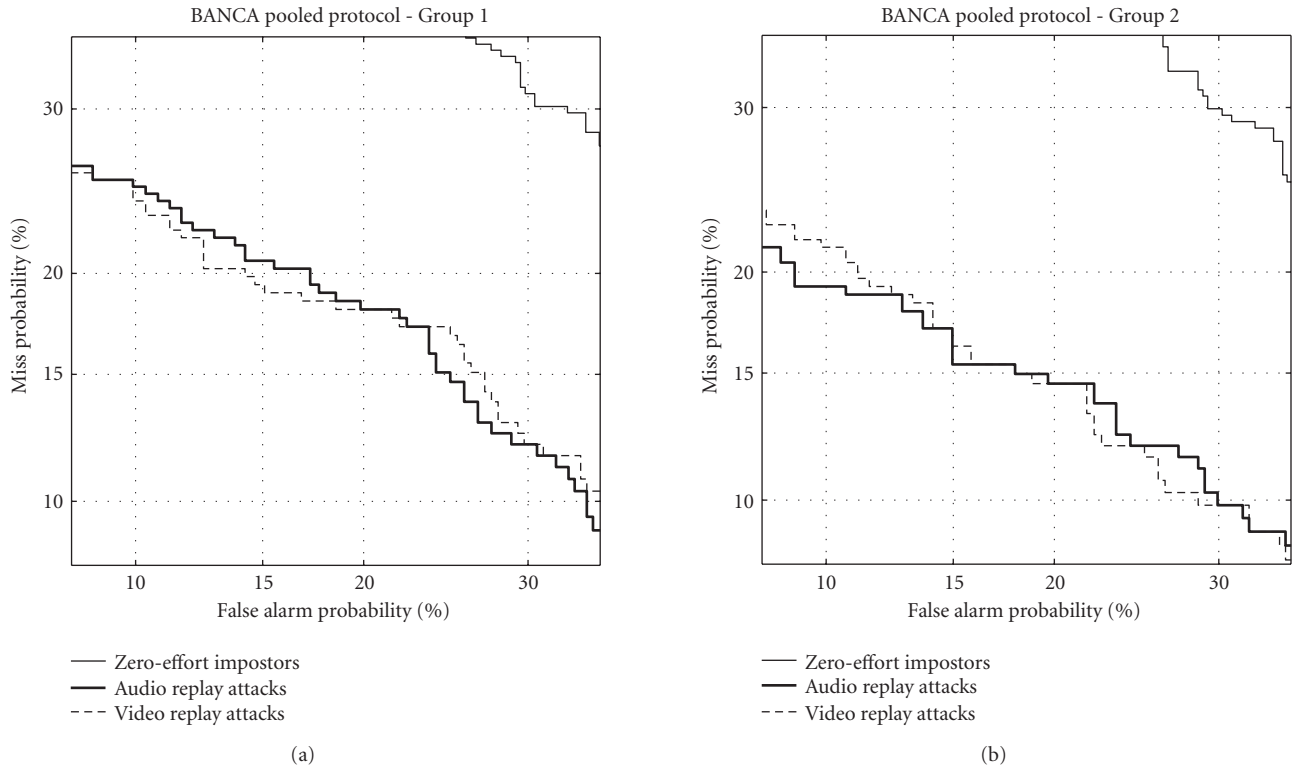

Figure 4: Identity verification with speech synchrony.

And also

In [31], audiovisual speech correspondence is used as a way of improving an algorithm for speech separation. The authors of [30] design filters for noise reduction, with the help of audiovisual speech correspondence.

7. CONCLUSION

This paper has reviewed techniques proposed in the literature to measure the degree of correspondence between audio and visual speech. However, it is very difficult to compare these methods since no common framework is shared among the laboratories working in this area. There was a monologue detection task (where using audiovisual speech correspondence showed to improve performance in [12]) in TRECVid 2002 but unfortunately it disappeared in the following sessions (2003 to 2006). Moreover, tests are often performed on very small datasets, sometimes only made of a couple of videos and difficult to reproduce. Therefore, drawing any conclusions about performance is not an easy task, the area covered in this review clearly lacks a common evaluation framework.

Nevertheless, experimental protocols and databases do exist for research in biometric authentication based on talking faces. We have therefore used the BANCA database and its predefined *Pooled* protocol to evaluate the performance of synchrony measures for biometrics, an EER of 32% was reached. The fact that this new modality is very little correlated to speaker verification and face recognition might also lead to significant improvement in a multimodal system based on the fusion of the three modalities [41].

ACKNOWLEDGMENT

The research leading to this paper was supported by the European Commission under Contract FP6-027026, Knowledge Space of semantic inference for automatic annotation and retrieval of multimedia content—K-Space.

REFERENCES

[1] G. Potamianos, C. Neti, J. Luettin, and I. Matthews, "Audio-visual automatic speech recognition: an overview," in *Issues in Visual and Audio-Visual Speech Processing*, G. Bailly, E. Vatikiotis-Bateson, and P. Perrier, Eds., chapter 10, MIT Press, Cambridge, Mass, USA, 2004.

[2] T. Chen, "Audiovisual speech processing," *IEEE Signal Processing Magazine*, vol. 18, no. 1, pp. 9–21, 2001.

[3] C. C. Chibelushi, F. Deravi, and J. S. Mason, "A review of speech-based bimodal recognition," *IEEE Transactions on Multimedia*, vol. 4, no. 1, pp. 23–37, 2002.

[4] J. P. Barker and F. Berthommier, "Evidence of correlation between acoustic and visual features of speech," in *Proceedings of the 14th International Congress of Phonetic Sciences (ICPhS '99)*, pp. 199–202, San Francisco, Calif, USA, August 1999.

[5] H. Yehia, P. Rubin, and E. Vatikiotis-Bateson, "Quantitative association of vocal-tract and facial behavior," *Speech Communication*, vol. 26, no. 1-2, pp. 23–43, 1998.

[6] E. Bailly-Baillière, S. Bengio, F. Bimbot, et al., "The BANCA database and evaluation protocol," in *Proceedings of the 4th International Conference on Audioand Video-Based Biometric Person Authentication (AVBPA '03)*, vol. 2688 of *Lecture Notes in Computer Science*, pp. 625–638, Springer, Guildford, UK, January 2003.

[7] J. Hershey and J. Movellan, "Audio-vision: using audio-visual synchrony to locate sounds," in *Advances in Neural Information Processing Systems 11*, M. S. Kearns, S. A. Solla, and D. A. Cohn, Eds., pp. 813–819, MIT Press, Cambridge, Mass, USA, 1999.

[8] H. Bredin, A. Miguel, I. H. Witten, and G. Chollet, "Detecting replay attacks in audiovisual identity verification," in *Proceedings of the 31st IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '06)*, vol. 1, pp. 621–624, Toulous, France, May 2006.

[9] J. W. Fisher III and T. Darrell, "Speaker association with signal-level audiovisual fusion," *IEEE Transactions on Multimedia*, vol. 6, no. 3, pp. 406–413, 2004.

[10] G. Chetty and M. Wagner, ""Liveness" verification in audio-video authentication," in *Proceedings of the 10th Australian International Conference on Speech Science and Technology (SST '04)*, pp. 358–363, Sydney, Australia, December 2004.

[11] R. Cutler and L. Davis, "Look who's talking: speaker detection using video and audio correlation," in *Proceedings of IEEE International Conference on Multimedia and Expo (ICME '00)*, vol. 3, pp. 1589–1592, New York, NY, USA, July-August 2000.

[12] G. Iyengar, H. J. Nock, and C. Neti, "Audio-visual synchrony for detection of monologues in video archives," in *Proceedings of IEEE International Conference on Multimedia and Expo (ICME '03)*, vol. 1, pp. 329–332, Baltimore, Md, USA, July 2003.

[13] H. J. Nock, G. Iyengar, and C. Neti, "Assessing face and speech consistency for monologue detection in video," in *Proceedings of the 10th ACM international Conference on Multimedia (MULTIMEDIA '02)*, pp. 303–306, Juan-les-Pins, France, December 2002.

[14] M. Slaney and M. Covell, "FaceSync: a linear operator for measuring synchronization of video facial images and audio tracks," in *Advances in Neural Information Processing Systems 13*, pp. 814–820, MIT Press, Cambridge, Mass, USA, 2000.

[15] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted Gaussian mixture models," *Digital Signal Processing*, vol. 10, no. 1–3, pp. 19–41, 2000.

[16] N. Sugamura and F. Itakura, "Speech analysis and synthesis methods developed at ECL in NTT-from LPC to LSP," *Speech Communications*, vol. 5, no. 2, pp. 199–215, 1986.

[17] C. Bregler and Y. Konig, ""Eigenlips" for robust speech recognition," in *Proceedings of the 19th IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '94)*, vol. 2, pp. 669–672, Adelaide, Australia, April 1994.

[18] M. Turk and A. Pentland, "Eigenfaces for recognition," *Journal of Cognitive Neuroscience*, vol. 3, no. 1, pp. 71–86, 1991.

[19] R. Goecke and B. Millar, "Statistical analysis of the relationship between audio and video speech parameters for Australian English," in *Proceedings of the ISCA Tutorial and Research Workshop on Audio Visual Speech Processing (AVSP '03)*, pp. 133–138, Saint-Jorioz, France, September 2003.

[20] N. Eveno and L. Besacier, "A speaker independent "liveness" test for audio-visual biometrics," in *Proceedings of the 9th European Conference on Speech Communication and Technology (EuroSpeech '05)*, pp. 3081–3084, Lisbon, Portugal, September 2005.

[21] N. Eveno and L. Besacier, "Co-inertia analysis for "liveness" test in audio-visual biometrics," in *Proceedings of the 4th International Symposium on Image and Signal Processing and Analysis (ISPA '05)*, pp. 257–261, Zagreb, Croatia, September 2005.

[22] N. Fox and R. B. Reilly, "Audio-visual speaker identification based on the use of dynamic audio and visual features," in *Proceedings of the 4th International Conference on Audio- and Video-Based Biometric Person Authentication (AVBPA '03)*, vol. 2688 of *Lecture Notes in Computer Science*, pp. 743–751, Springer, Guildford, UK, June 2003.

[23] C. C. Chibelushi, J. S. Mason, and F. Deravi, "Integrated person identification using voice and facial features," in *IEE Colloquium on Image Processing for Security Applications*, vol. 4, pp. 1–5, London, UK, March 1997.

[24] A. Hyvärinen, "Survey on independent component analysis," *Neural Computing Surveys*, vol. 2, pp. 94–128, 1999.

[25] D. Sodoyer, L. Girin, C. Jutten, and J.-L. Schwartz, "Speech extraction based on ICA and audio-visual coherence," in *Proceedings of the 7th International Symposium on Signal Processing and Its Applications (ISSPA '03)*, vol. 2, pp. 65–68, Paris, France, July 2003.

[26] P. Smaragdis and M. Casey, "Audio/visual independent components," in *Proceedings of the 4th International Symposium on Independent Component Analysis and Blind Signal Separation (ICA '03)*, pp. 709–714, Nara, Japan, April 2003.

[27] ICA, http://www.cis.hut.fi/projects/ica/fastica/.

[28] Canonical Correlation Analysis. http://people.imt.liu.se/~magnus/cca/.

[29] S. Dolédec and D. Chessel, "Co-inertia analysis: an alternative method for studying species-environment relationships," *Freshwater Biology*, vol. 31, pp. 277–294, 1994.

[30] J. W. Fisher, T. Darrell, W. T. Freeman, and P. Viola, "Learning joint statistical models for audio-visual fusion and segregation," in *Advances in Neural Information Processing Systems 13*, T. K. Leen, T. G. Dietterich, and V. Tresp, Eds., pp. 772–778, MIT Press, Cambridge, Mass, USA, 2001.

[31] D. Sodoyer, J.-L. Schwartz, L. Girin, J. Klinkisch, and C. Jutten, "Separation of audio-visual speech sources: a new approach exploiting the audio-visual coherence of speech stimuli," *EURASIP Journal on Applied Signal Processing*, vol. 2002, no. 11, pp. 1165–1173, 2002.

[32] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, 1989.

[33] S. Bengio, "An asynchronous hidden Markov model for audio-visual speech recognition," in *Advances in Neural Information Processing Systems 15*, S. Becker, S. Thrun, and K. Obermayer, Eds., pp. 1213–1220, MIT Press, Cambridge, Mass, USA, 2003.

[34] H. Bredin, G. Aversano, C. Mokbel, and G. Chollet, "The biosecure talking-face reference system," in *Proceedings of the 2nd Workshop on Multimodal User Authentication (MMUA '06)*, Toulouse, France, May 2006.

[35] K. Messer, J. Matas, J. Kittler, J. Luettin, and G. Maitre, "XM2VTSDB: the extended M2VTS database," in *Proceedings of International Conference on Audio- and Video-Based Biometric Person Authentication (AVBPA '99)*, pp. 72–77, Washington, DC, USA, March 1999.

[36] BT-DAVID, http://eegalilee.swan.ac.uk/.

[37] S. Garcia-Salicetti, C. Beumier, G. Chollet, et al., "BIOMET: a multimodal person authentication database including face, voice, fingerprint, hand and signature modalities," in *Proceedings of the 4th International Conference on Audio-and Video-Based Biometric Person Authentication (AVBPA '03)*, pp. 845–853, Guildford, UK, June 2003.

[38] A. F. Martin, G. R. Doddington, T. Kamm, M. Ordowski, and M. Przybocki, "The DET curve in assessment of detection task performance," in *Proceedings of the 5th European Conference on Speech Communication and Technology (EuroSpeech '97)*, vol. 4, pp. 1895–1898, Rhodes, Greece, September 1997.

[39] M. E. Sargin, E. Erzin, Y. Yemez, and A. M. Tekalp, "Multimodal speaker identification using canonical correlation analysis," in *Proceedings of the 31st IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '06)*, vol. 1, pp. 613–616, Toulouse, France, May 2006.

[40] Text Retrieval Conference Video Track. http://trec.nist.gov/.

[41] H. Bredin and G. Chollet, "Audio-visual speech synchrony measure for talking-face identity verification," in *Proceedings of the 32nd IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '07)*, Honolulu, Hawaii, USA, April 2007.

Hervé Bredin is a French Ph.D. candidate at the Signal and Image Processing Department (http://www.tsi.enst.fr/~bredin) of the "Grande École" Télécom Paris, (http//www.enst.fr) from which he received his Engineering diploma in 2004, focusing mostly on signal and image processing, pattern recognition, and human-computer interactions. His research deals with biometrics and, more precisely, audiovisual identity verification based on talking faces and its robustness to high-effort forgery (such as replay attacks, face animation, or voice transformation).

Gérard Chollet's education was centered on mathematics (DUES-MP), physics (Ma"trise), engineering and computer sciences (DEA). He studied linguistics, electrical engineering, and computer science at the University of California, Santa Barbara, where he was granted the Ph.D. degree in computer science and linguistics. He taught courses in phonetics, speech processing, and Psycholinguistic in the Speech and Hearing Department at Memphis State University in 1976-1977. Then, he had a dual affiliation with the Computer Science and Speech Departments at the University of Florida in 1977-1978. He joined CNRS (the french public research agency) in 1978 at the Institut de Phonétique in Aix en Provence. In 1981, he was asked to take in charge of the Speech Research Group of Alcatel. In 1983, he joined a newly created CNRS research unit at ENST where he was Head of the Speech Group. In 1992, he participated in the development of IDIAP, a new research laboratory of the "Fondation Dalle Molle" in Martigny, Switzerland. Since 1996, he is back, full time 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.